Ghent University

Faculty of Sciences

Department of Plant Biotechnology and Bioinformatics

Post-translational regulation and the evolution of eukaryotic genomes

This thesis is submitted as fulfillment of the requirements for the degree of

PhD in Sciences, Bioinformatics

20th of December 2011

YING HE

Promotor: **Prof. Dr. Yves Van de Peer**

VIB / Plant Systems Biology

Bioinformatics and Systems Biology Group

Technologiepark 927, B-9000 Gent, Belgium

Examination Committee

Prof. Dr. Geert De Jaeger (Chair)

Department of Plant Biotechnology and Bioinformatics, Ghent University

Flanders Institute for Biotechnology (VIB)

Prof. Dr. Yves Van de Peer (Promotor)

Department of Plant Biotechnology and Bioinformatics, Ghent University

Flanders Institute for Biotechnology (VIB)

Dr. Grigoris Amoutzias

Department of Biochemistry & Biotechnology

University of Thessaly, Greece

Prof. Dr.Jack Leunissen

Lab. Bioinformatics,

Wageningen University, the Netherlands

Prof. Dr. Klaas Vandepoele

Department of Plant Biotechnology and Bioinformatics, Ghent University

Flanders Institute for Biotechnology (VIB)

Dr. Stefanie De Bodt

Department of Plant Biotechnology and Bioinformatics, Ghent University

Flanders Institute for Biotechnology (VIB)

Dr. Pieter De Bleser

Department for Molecular Biomedical Research (DMBR), Ghent University

Flanders Institute for Biotechnology (VIB)

Acknowledgements

I would like to start by thanking those who are the first ones to read this thesis, my jury members. I greatly appreciate all your comments that have helped improve this manuscript. And then I would like to thank the person who offered me the chance to be a bioinformatician. Four years ago, when I was quite bored with my life in the Netherlands, I got an opportunity for a PhD student interview here in Gent. Yves together with many other group members welcomed me with a very delicious lunch, which impressed me a big deal both in research and life style. Yves, I deeply thank you for providing this great opportunity to work in your lab and being always supportive in every single way. As a PhD student, I have been involving in several very interesting and exciting projects and had access to all kinds of resources. I would say I have really enjoyed my work and stay here in such a famous and successful group, so thank you for guiding and trusting me during these four years. The annual ski trip is awesome! Although people will say it is crazy to go to black slope the first time of skiing, I am glad that we both think it is not.

I would also want to thank Greg for giving me a wealth of advice and being not just a colleague, a collaborator but a truly good friend. I really appreciate all your help and patience and I am really happy to know you and work closely with you. I thank all my colleagues in the bioinformatics group for helping me solve many technical problems and being tolerant when I tried (not intentionally) to crash midas again and again. I want to give some special thanks to people in the small sysbiol room who have been friendly and super nice to me since the first day I joined the group. Anagha, (my Indian sister) is just there for me whenever and wherever needed; Eric, Tom and Vanessa, you are the best officemates ever. Also for all the people (especially the girls, Cindy, Elisabeth, Litsa, Sandra, Sara) I have been hanging out with for dinners, movies and funs, thank you for keeping me companies and sharing all ups and downs in life. I also want to thank all my Chinese friends for the nice lunch time

together and being supportive for each other (XFF and XYY especially). And also to many other people that have made my everyday life unforgettable. Last but not the least, I want to thank my family members: 感谢我老姐我老弟在我不在国内的时候对我父母的照顾; 感谢我伯父伯母对我的支持和鼓励; 最要感谢的就是我的父母, 对我无条件的信任和爱. 有你们,我很幸运.

Scope

Organismal complexity is the result of much more than the number of nucleotides that a genome is composed of and the number of protein-coding genes in that genome. What is the size of the genome and how many genes are there? The answer to this question surprisingly reveals little about organism's complexity. Organismal complexity does not rely solely on the number of building blocks (genes, regulatory elements), but also on how these blocks are combined (Xia, Fu et al. 2008). As the complexity of organisms (measured by the number of cell types) increases, the collection of transcription factors expands. The complexity of the protein interaction network and regulatory network may provide additional explanations to the lack of correlation between the complexity of an organism and its DNA content.

The regulation of eukaryotic gene expression occurs at the levels of transcription, RNA splicing, RNA stability, translation and post-translational modification. Different types of molecular networks (*e.g.* transcriptional, phosphorylation, genetic interaction, miRNA, protein-protein interaction and metabolic pathway networks) rewire at different rates during evolution. The top two fastest evolving regulatory networks, transcription factor-target regulatory networks and kinase-substrate phosphorylation networks are of great importance to regulation in species evolution (Shou, Bhardwaj et al. 2011).

The structure of the genomes that we observe today is the result of billions of years of evolution. Gene and genome duplications create novel genetic material on which evolution can work and have therefore been recognized as a major source of innovation for many eukaryotic lineages (Ohno 1970). Following duplication, the most likely fate is gene loss; however, a considerable fraction of duplicated genes survive

(either by sub-functionalization and/or by neo-functionalization). The loss and gain of interactions in gene regulation will contribute to the rewiring and adaptation of regulatory networks. After duplications, not all genes have the same probability of survival and it is not fully understood what evolutionary forces determine the pattern of gene retention.

Genes are merely the recipe, while proteins are the functional units and the machine that drives much of biology. It is not enough to know if the protein is going to be present in a given condition, but also the events that modify the function of this protein, namely post-translational modifications. Phosphorylation is considered to be the most abundant modification. A large number of phosphoproteomics experiments have been performed with *Saccharomyces cerevisiae.* Thus it has provided a great opportunity to explore these datasets and to study the properties of the yeast phosphoproteome (Chapter 2). We have investigated the impact of post-translational modifications, in particular phosphorylation, on the evolution of the genome (Chapter 3) as well as the substitution pattern of phosphorylation sites (Chapter 4).

General properties of the yeast phosphoproteome, the impact of post-translational modification (in particular phosphorylation) on genome evolution and the evolutionary pattern of phosphorylation sites will be investigated and discussed in this thesis. Better understanding of gene regulation (particularly at the post-translational level) and genome evolution will improve our knowledge concerning the organismal complexity of eukaryotic species.

Summary

With the advent of phosphoproteomics in the past decade, high-throughput detection of phosphorylation sites, at the genome scale, became available. Therefore, it provides a great opportunity to study the general properties of the phosphoproteome and the impact of protein phosphorylation, probably the most abundant post-translational modifications, on the evolution of the genome, by using computational and evolutionary analyses.

A large number of phosphoproteomics experiments have been performed with *Saccharomyces cerevisiae*, under a reasonably wide range of conditions. With a wealth of relevant functional genomic information available for this organism, all of these factors should assist in an in-depth bioinformatics analysis of the yeast phosphoproteome. Observations from the yeast phosphoproteome may help draw general conclusions and formulate new questions about phosphorylation and understand certain biological processes in other eukaryotes.

In Chapter 2, general properties of the yeast phosphoproteome have been investigated. A high-quality curated phosphoproteomic dataset is assembled from 12 publicly available phosphoproteomic datasets. Recently, concerns have been raised about the possibility of detecting non-functional phosphorylation sites that are technical false-positives or of low-stoichiometry within the cell. The HQ dataset we have provided in Chapter 2 is considered to be filtered of such noise and has been used to study the general properties of the yeast phosphoproteome in a comprehensive way.

The impact of post-translational regulation, in particular phosphorylation, on eukaryotic genome evolution has been studied in Chapter 3. Interestingly, we

reported for the first time in literature that phosphorylation affects the retention of duplicated genes (especially after genome duplication) in the fungal lineages. This finding may help better understand the basic mechanisms of gene retention and genome evolution.

In addition, the general properties (amino acid substitution) of phosphorylation sites have been investigated in Chapter 4. It has been observed in this study that phosphorylated serines, compared to their non-modified counterparts, tend to substitute more frequently to amino acids (aspartic acid, glutamic acid) that shared similar (phosphomimetic) properties and less frequently to alanine which resembles non-phosphorylation status.

Abbreviations

A	Alanine
AA	Amino Acid
CDS	Coding sequence
D	Aspartic acid
DNA	Deoxyribonucleic acid
E	Glutamic acid
GO	Gene Ontology
HQ	High Quality
HTP	High-throughput
ID	Intrinsic Disorder
IMAC	Immobilized Metal-Affinity Chromatography
LTP	Low-throughput
mRNA	messenger RNA
MS	Mass Spectrometry
mya	million years ago
NGS	Next Generation Sequecing
ORF	Open Reading Frame
PO4	phosphate
PPI	Protein-Protein Interaction
p-sites	phosphorylation sites
PTM	Post-Translational Modification
RSS	Return to Single Status gene
S	Serine
SCX	Strong cation exchange chromatography
SILAC	Stable Isotope labeling with amino acids in cell culture
SSD	Small Scale gene Duplication
T	Threonine
TF	Transcription Factor
TiO2	titanium dioxide
WGD	Whole Genome Duplication
Y	Tyrosine
YGOB	Yeast Genome Order Browser
Y2H	Yeast Two Hybrid

Contents

1 Introduction

1.1 Preface

In the first part of this chapter I will present an introduction on the study of eukaryotic gene regulation and the levels where regulation occurs. In the second part I will outline the “how”s and “why”s to evaluate the properties of the budding yeast phosphoproteome and address the impact on the fate of duplicated genes of post-translational regulation.

1.2 Gene regulation in eukaryotes

1.2.1 Organismal complexity

A mystery has been surrounding the extensive variations in genome size among eukaryotic species, which is more commonly known as the C-value paradox. This paradox is characterized by the observation that genome size (measured by the amount of DNA contents found in a haploid genome) does not have a correlation with organismal complexity (Gregory 2001). For instance, the genome size of the unicellular protist Amoeba is roughly 200 times larger than that of human, which is due to the collection of huge amounts of non-coding DNAs in Amoeba. Also, having more protein-coding genes does not necessarily translate into greater complexity which sometimes is addressed as N-value paradox (Pray 2008). Organismal complexity is the result of much more than the number of nucleotides that a genome is composed of and the number of protein-coding genes in that genome. It is not the

size of the genome or the number of genes that matters, but how these genes are regulated.

Organisms do not try to generate complexity, which is the by-product of adaption, survival and evolution. Some parasitic organisms become less complex because they do not need it. Much of organismal complexity derives from how various genes are expressed and combined (Pray 2008). Organismal complexity does not rely solely on the number of building blocks (genes, regulatory elements), but on how these blocks are combined. For example, Myc, Mad and Max all belong to the basic helix-loop-helix leucine zipper (bHLHZ) family of transcription factors in human. bHLHZ segments of these three TFs allow the formation of Myc-Max and Mad-Max heterodimers and the Max-Max homodimer that are responsible for distinct functions and roles in the regulation of transcription (Nair and Burley 2003).

It has been reported that as the complexity of organisms (measured by the number of cell types) increases, the collection of transcription factors in metazoa expands (Hedges, Blair et al. 2004; Vogel and Chothia 2006; Amoutzias, Veron et al. 2007). Expansions of transcription-associated proteins are also reported to be directly correlated to the morphological complexity of an organism (Lang, Weiche et al. 2010). Protein domains with signaling and regulatory functions have been repeatedly found highly expanded in various metazoan proteomes and that protein-protein interaction (PPI) domains specific to or expanded in metazoa preferentially participate in signaling and regulation (Xia, Fu et al. 2008). Since transient interactions (transcription factor-target binding, phosphorylation) dominate gene regulation and signal transduction and given the established link between organismal complexity (at least in terms of distinct cell types) and an increase in the percentage of signal transducers and TFs within a genome (Ranea et al., 2005; van Nimwegen, 2003), it is tempting to assume that whole genome duplication (WGD) is one of the major contributors of raw genetic material to increase biological complexity. Previous work has shown that transcription factors, signal transducers, and developmental genes have been preferentially retained after WGD events (Davis and Petrov 2005; Maere,

De Bodt et al. 2005; Freeling and Thomas 2006). The complexity of the protein interaction network and regulatory network may provide additional explanations to the lack of correlation between the complexity of an organism and its DNA content.

1.2.2 Different levels of gene regulation

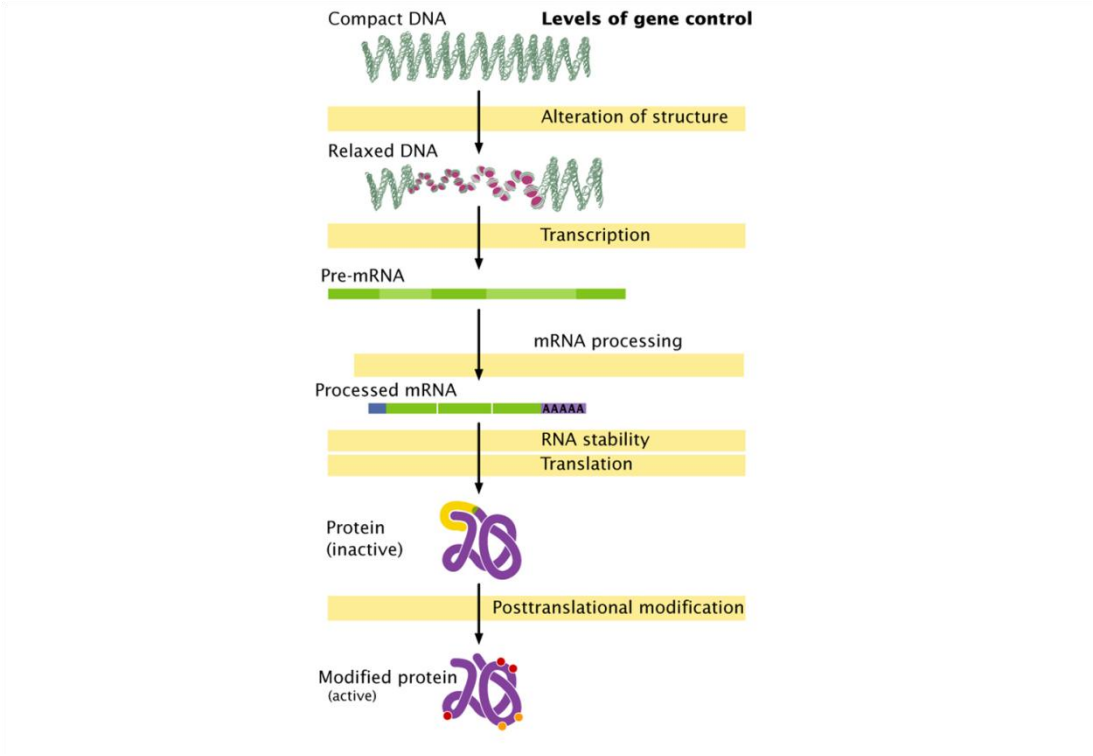


Figure 1.1: Different stages in regulation of eukaryotic gene expression are shown in the figure above from transcription, RNA splicing, RNA stability, translation to post-translational modification. Gene expression may be modulated at any step, from DNA-RNA transcription to the post-translational modification of a protein. In eukaryotes, the accessibility of large regions of DNA can depend on its chromatin structure and be altered by modifications such as DNA methylation, ncRNA, or DNA-binding protein, which may up or down regulate gene expression. After the DNA is transcribed, a primary transcript of RNA (pre-mRNA) is formed. mRNA processing which refers to a series of modifications on pre-mRNA will lead to a mature mRNA. In eukaryotes RNA is stabilized by certain post-transcriptional modifications, particularly the 5' cap and poly-adenylated tail. Each protein exists as an unfolded polypeptide or random coil when translated from a sequence of mRNA to a linear chain of amino acids. Post-translational modification is the chemical modification of a protein after its translation. Proteins are activated after post-translational modification by having structural changes, being attached by other biochemical functional groups (taken and modified from Fig16-1, Genetics, Second Edition, 2005 W.H. Freeman and Company).

Genomes are more than linear sequences and knowing the sequences of a genome is insufficient to understand its physiological function (Misteli 2007). Regulation of gene expression refers to the control of the amount and timing of appearance of the

functional product of a gene (Figure 1.1). From a finite set of genes, many different phenotypes can emerge, depending on their spatiotemporal expression, the splicing of their transcripts, and the post-translational modifications of their protein products (Reik 2007).

The regulation of eukaryotic gene expression occurs at the levels of transcription, RNA splicing, RNA stability, translation and post-translational modification (PTM) (see Figure 1.1). It has been demonstrated that different types of molecular networks rewire at different rates with their order being (from fast to slow) transcriptional, phosphorylation, genetic interaction, miRNA, protein-protein interaction and metabolic pathway networks (Shou, Bhardwaj et al. 2011).

1.2.3 Transcription factors and transcriptional regulation

Undoubtedly, transcription factors (TFs), as controllers of transcription initiation, influence such important biological functions as gene regulation and the complexity of development and differentiation. Therefore, it is not surprising that organismal complexity (measured by the number of distinct cell types) correlates with an expansion of certain gene categories that are involved in gene regulation and signal transduction (Miyata and Suga 2001; Amoutzias, Robertson et al. 2004). There also seems to be a correlation between the number and coverage of protein-protein interaction (PPI) domains per protein (Xia, Fu et al. 2008).

DNA-binding and activation domains are the two common domains that are present in many TF proteins. Other optional domains can be found, such as dimerization, ligand-binding and repressor domain (see Figure 1.2) (Latchman 2001):

- DNA-Binding Domain (DBD): ‘helix-turn-helix’ and ‘zinc finger’ are the two most common motifs of DBD, TFs are classified based on their DBD;

- Transcription activation domain (TAD): which contains interaction surfaces for other proteins such as transcription co-regulators (Warnmark, Treuter et al. 2003), ‘Gal4’ for example.

- Signal sensing domain (SSD): an optional domain that senses external signals and, in response, these signals are transmitted and gene expression is up- or down-regulated accordingly; including dimerization domain such as ‘leucine zipper’, ‘helix-loop-helix’ or a ligand binding domain.

Figure 1.2: Schematic diagram of the amino acid sequence (amino terminus to the left and carboxylic acid terminus to the right) of a prototypical transcription factor that contains (1) a DNA-binding domain (DBD), (2) signal-sensing domain (SSD), and a trans-activation domain (TAD). The order of placement and the number of domains may differ in various types of transcription factors. In addition, the trans-activation and signal-sensing functions are frequently contained within the same domain (taken from wikipedia).

Various segments of a TF, like the DNA-binding domain, dimerization, activation, ligand-binding or other domains that are involved in protein-protein interactions are conserved at different levels (Hsia and McGinnis 2003).

1.2.4 Proteins and post-translational regulation

Translation is also a part of the overall process of gene expression. After transcription is initiated, DNA is transcribed in the nucleus into messenger RNA (mRNA). Since protein translation occurs in the ribosome, mRNA has to be exported to the cytoplasm first. Amino acids are delivered by transfer RNAs that are complementary to each base pair triplet codon in the messenger RNA produced by transcription. These amino acids are then bonded together to form a specific chain or polypeptide,

which will later fold into an active protein (see Figure 1.1) (Berg, Tymoczko et al. 2002).

After translation, the post-translational modification (PTM) of amino acids extends the range of functions of the protein by attaching it to other biochemical functional groups such as acetate, phosphate, various lipids and carbohydrates, by changing the chemical nature of an amino acid (e.g. citrullination) or by making structural changes, like the formation of disulfide bridges. Post-translational regulation refers to the control of the levels of active proteins (Lodish, Berk et al. 2000). One of the most common posttranslational modifications, phosphorylation, will be discussed in more detail in the following sections (see section 1.3).

1.3 Exploring the yeast phosphoproteome

1.3.1 The structure of protein

The ability to fold into a certain structure is one of the most distinguishing features of polypeptides. Proteins can change their conformations from one to another, which is often associated with a signaling event. Therefore, the structure of a protein serves as a medium and the function of a protein or the activity of an enzyme is regulated accordingly (Lodish, Berk et al. 2000). However, a stable or folded conformation is not a must for all proteins in order to carry out their specific biological functions. Certain regions within proteins, or in some cases even the entire proteins, are not ordered into a unique tertiary structure, but instead appear to exist as ensembles of structures (see Figure 1.3) (Dunker and Obradovic 2001). A high percentage of cell-signaling proteins are predicted to have long disordered regions, which may indicate the general importance of intrinsic disorder for signaling and regulation (Iakoucheva, Brown et al. 2002). A previous study has reported that proteins with intrinsically

disordered regions are often involved in protein modifications including phosphorylations (Dunker, Brown et al. 2002). The analyses of yeast protein structure and its function (in terms of phosphorylation) are presented in Chapter 2 and 3 in detail followed by a series of discussions that are consistent with previous researches.

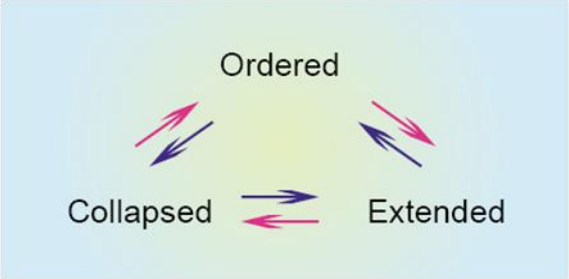


Figure 1.3: The protein trinity. Native proteins may exist in one of three states: ordered, collapsed-disordered, or extended-disordered (taken from Dunker and Obradovic 2001).

1.3.2 Proteomics

Genomics provides sequence information of the full complement of genetic materials in an organism. Transcriptomics, which refers to gene expression measured by transcriptional regulation of genes, is investigated by using DNA microarray and RNA-sequencing technologies (Duggan, Bittner et al. 1999; Ozsolak and Milos 2011) .

The third well known ‘omics’ science is proteomics. A proteome is the complete set of proteins expressed by an organism at a specific time/condition in a particular cell or tissue type. Unlike the genome, which is characterized by its stability, proteome actively changes in response to various factors, including the organism's developmental stage and both internal and external conditions (Brown 2002). Proteomics is the study of the proteome to understand where/when proteins are expressed and interact with one another. Genomics measures the genotype of an

organism and uses fast developing technologies for decoding nucleic acid sequences, whereas proteomics measures the phenotype, shaped by both the genotype and the past and present environment of the organism, and uses mass spectrometry (MS)-based approaches (Yates, Ruse et al. 2009; Cox and Mann 2010). Proteins undergo modifications, which may occur either before or after translation, and proteomics can directly and quantitatively determine the modification state of proteins. Phosphorylation is a key reversible post-translational modification that regulates protein function, subcellular localization, complex formation, degradation of proteins and therefore cell signaling networks (Lodish, Berk et al. 2000). In this thesis, some interesting facts are revealed with use of yeast phosphoproteomics from several high-throughput experiments (HTPs) in chapters 2, 3 and 4.

1.3.3 Protein phosphorylation

In many cases, the result of a signaling pathway is the post-translational modification of target-cell proteins whose activities have been changed (from inactive to active or the opposite, shown in the last step in Figure 1.1). Protein phosphorylation may be the most common post-translational modification, which plays a significant role in a wide range of cellular processes, such as gene regulation, cell cycle control during metabolic processes, and the organization of the cytoskeleton and cell adhesion in secretory processes (Barford, Das et al. 1998).

Phosphorylation is the addition of a phosphate (PO_4) group to a protein or other organic molecule (see Figure 1.4). Reversible phosphorylation of proteins is an important regulatory mechanism that occurs in both prokaryotic and eukaryotic organisms, where many protein enzymes and receptors are activated or deactivated by phosphorylation and dephosphorylation (Chang and Stewart 1998). Enzymes called kinases and phosphatases, which are respectively responsible for phosphorylation and dephosphorylation, are involved in the reversible process of

protein phosphorylation. Phosphorylation often occurs at multiple residues within a protein and in most cases by different protein kinases (Berg, Tymoczko et al. 2002; Ubersax and Ferrell 2007). The result of phosphorylation is an allosteric conformation change in the structure of the protein with its function being switched 'on' or 'off' accordingly. During phosphorylation, protein is regulated by covalent attachment of phosphate, in ester linkage, to the side-chain hydroxyl group of a particular amino acid residue (serine, threonine or tyrosine, the three dominant phosphorylation sites in eukaryotes, see Figure 1.4) (Thingholm, Jensen et al. 2009).

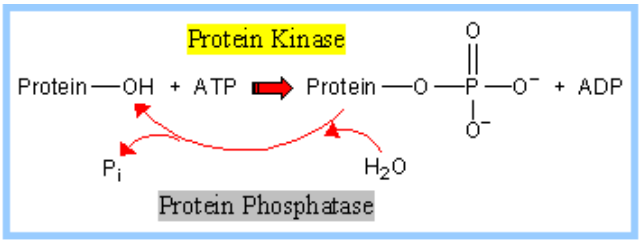


Figure 1.4: A protein kinase transfers the terminal phosphate of ATP to a hydroxyl group on a protein; a protein phosphatase catalyzes removal of the phosphate by hydrolysis (taken from Joyce J. Diwan's online course on Biochemistry of Metabolism).

1.3.4 Phosphoproteome techniques

Phosphoproteomics refers to a large-scale analysis of protein phosphorylation (in response to various stimuli or developmental states) using mass spectrometry (MS)-based strategies (Johnson and Hunter 2004). Recently, several yeast phosphoproteomic analyses have been reported (Gruhler, Olsen et al. 2005; Chi, Huttenhower et al. 2007; Li, Gerber et al. 2007; Albuquerque, Smolka et al. 2008; Bodenmiller, Campbell et al. 2008; Beltrao, Trinidad et al. 2009; Gnad, de Godoy et al. 2009; Holt, Tuch et al. 2009; Huber, Bodenmiller et al. 2009; Soufi, Kelstrup et al. 2009; Stark, Su et al. 2010). Phosphorylation is a transient modification that leads to changes in the conformation, activity, and interactions of a protein within a very short time frame, therefore phosphoproteins are often of very low abundance (Thingholm, Jensen et al. 2009). One major challenge that phosphoproteomics faces is the enrichment of phosphopeptides from these low abundance proteins (Gruhler, Olsen

et al. 2005). Many phosphoproteomics studies (Albuquerque, Smolka et al. 2008) use immobilized metal-affinity chromatography (IMAC) to enrich the sample for phosphopeptide by which either Ga (III) or Fe (III) is used as ligand. Strong cation exchange chromatography (SCX) is also widely used to fractionate peptides (Taouatas, Altelaar et al. 2009) (see Figure 1.5).

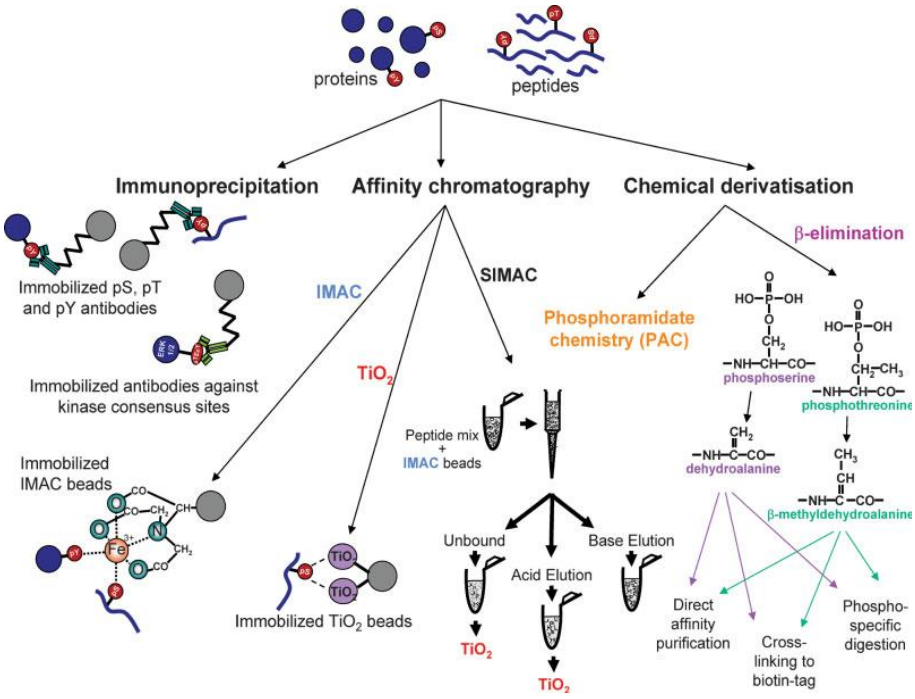


Figure 1.5: Strategies for phospho-specific enrichment. Most commonly used strategies for immunoprecipitation, affinity chromatography or chemical modification applied for enrichment of phosphoproteins and phosphopeptides are illustrated (taken from (Thingholm, Jensen et al. 2009)).

Recent success in the large-scale exploration of the yeast phosphoproteome owes much to these enrichment techniques (IMAC, SCX or the combination of two) as well as the development of MS instrumentation such as the linear ion trap and orbitrap (Gruhler, Olsen et al. 2005; Li, Gerber et al. 2007; Trinidad, Thalhammer et al. 2008). Stable isotope labeling with amino acids in cell culture (SILAC) is a simple and straightforward approach for *in vivo* incorporation of a label into proteins for MS-based quantitative proteomics. It has been applied to two cell populations with 'light' or 'heavy' form of both arginine and lysine of proteins in previous phosphoproteomics studies (Ong, Blagoev et al. 2002; de Godoy, Olsen et al. 2008). In the same two

experiments, titanium dioxide chromatography (TiO_2) is applied for phosphorylated peptide enrichment after lysis, 1:1 mixing and trypsin digestion. Identification of phosphopeptides with SILAC is very accurate (de Godoy, Olsen et al. 2008).

1.3.5 Identification of phosphopeptides and phosphorylation sites

Many approaches and algorithms have been applied to peptide and protein identification by searching a sequence database using data from MS experiments (Sadygov, Cociorva et al. 2004). Although difference is observed in detailed implementation from different reported methods, the general concept is similar: the experimental data are compared with peptide and peptide fragment mass values that are calculated on the basis of cleavage rules applied to the protein sequences in the database (Nesvizhskii 2007).

Identified peptide sequences are assigned to protein entries in the database afterwards, using a probability based scoring algorithm (MASCOT for example). By applying certain thresholds on scores, we can judge whether the result is statistically significant. The thresholds for different scores that we have used in our analysis will be presented in the following chapter (see Table 2.1 and Figure 1.6). There are some concerns about the correct placement of phosphorylation sites. Probability-based approaches have been developed to calculate the likelihood of matching given ions to specific phosphorylation site locations (Gnad, Ren et al. 2007). Cutoffs on different probability-based scores for proper placement of phosphorylation sites are applied accordingly (Table 2.1 and Figure 1.6). With these filters, the phosphoproteomics dataset we have generated is of high confidence, with cutoffs on experimental scores and without any ambiguous phosphorylation sites.

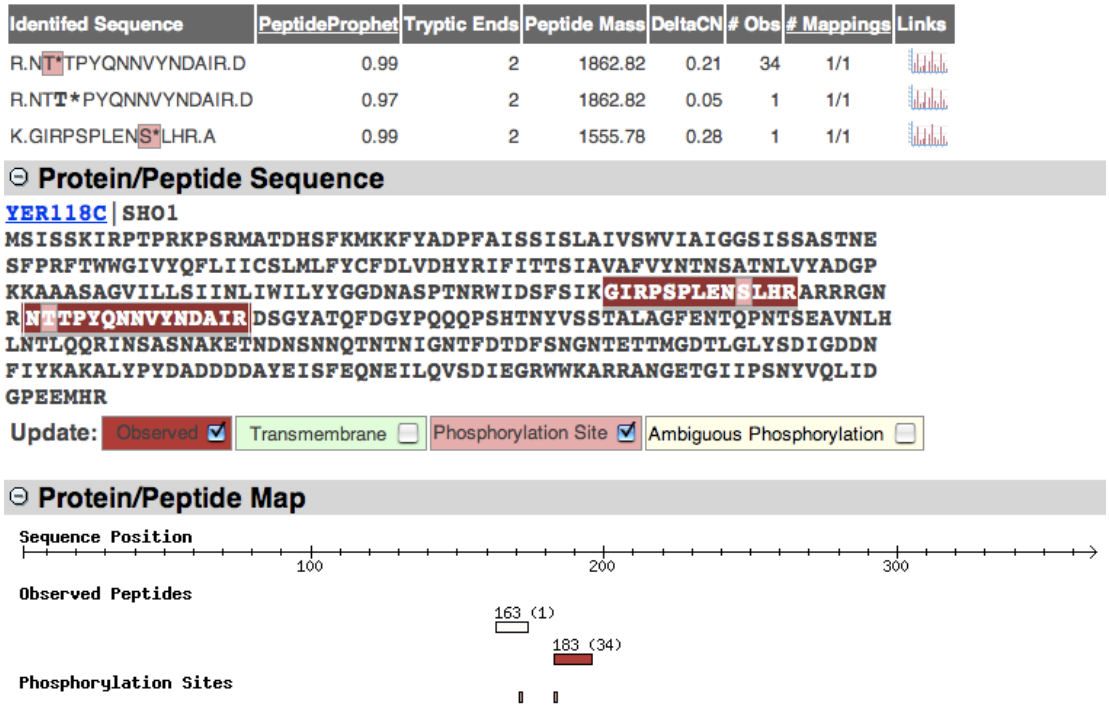


Figure 1.6: Two experimentally identified phosphopeptides with three potential phosphorylation sites are mapped to yeast protein YER118C. With the applied cutoff score, only two phosphorylation sites are assigned to this protein with high confidence (peptide search against Phosphopep2 database, (Bodenmiller, Campbell et al. 2008)).

1.3.6 The yeast phosphoproteome

The application of mass spectrometry combined with affinity techniques that are enriched for phosphopeptides has revolutionized the field of phosphoproteomics, such that hundreds, or even thousands of phosphorylation sites (p-sites) may be identified in a single experiment. However, as with any high throughput (HTP) technique, there are concerns about data quality and potential biases in the enrichment and identification procedures (Bodenmiller, Mueller et al. 2007; Lienhard 2008; Landry, Levy et al. 2009). Relatively low abundance of many phosphoproteins, low phosphorylation stoichiometry and the dynamic regulation of phosphoproteins put high demands on the analytical methods required for the studies of phosphoproteins. Sensitive, comprehensive and highly specific analytical strategies are used in order to characterize a meaningful number of modified proteins. Non-functional sites (off-target phosphorylation) are then identified due to high sensitivity of MS instruments

(Landry, Levy et al. 2009). Therefore, there is a need for stringent data evaluation to filter out possibly spurious p-sites before drawing any general conclusions about the structure and properties of a phosphoproteome.

A large proportion of intracellular proteins (up to 30%) is assumed to be phosphorylated at any given time, and it is perhaps not surprising that the human genome contains about 500 protein kinase genes, corresponding to about 2% of all human genes (Hunter 2000; Manning, Plowman et al. 2002). Phosphorylation usually occurs on serine, threonine, and tyrosine residues in eukaryotic proteins, which account for 86%, 12% and 2% respectively of the human phosphoproteome. Tens of thousands of phosphorylation sites are uncovered in the human proteome (Beausoleil, Jedrychowski et al. 2004; Olsen and Mann 2004; Amanchy, Kalume et al. 2005; Thelemann, Petti et al. 2005).

The budding yeast *Saccharomyces cerevisiae* is widely used as a model organism to gain biological insight in the basic functions of the eukaryotic cell. It also has an important role in industry for bread fermentation and ethanol (beer) production. Its genome was the first eukaryotic one to be completely sequenced (Goffeau, Barrell et al. 1996). There are several reasons for using yeast to benchmark novel phosphoproteomics technologies. The most important of which is that a large number of phosphoproteomics experiments have been performed with *Saccharomyces cerevisiae*, under a reasonably wide range of conditions (Gruhler, Olsen et al. 2005; Chi, Huttenhower et al. 2007; Li, Gerber et al. 2007; Albuquerque, Smolka et al. 2008; Bodenmiller, Campbell et al. 2008; Beltrao, Trinidad et al. 2009; Gnad, de Godoy et al. 2009; Holt, Tuch et al. 2009; Huber, Bodenmiller et al. 2009; Soufi, Kelstrup et al. 2009; Stark, Su et al. 2010). Moreover, a large fraction (~80%) of the predicted yeast phosphoproteome has been found to be expressed under normal laboratory growth conditions (Ghaemmaghami, Huh et al. 2003). Finally, there is a wealth of relevant functional genomic information available for the organism, including data on protein abundance, half-lives, and the number of kinases targeting a given protein

(Ghaemmaghami, Huh et al. 2003; Ptacek, Devgan et al. 2005; Belle, Tanay et al. 2006; Newman, Ghaemmaghami et al. 2006), amongst others. All of these factors should assist in an in-depth bioinformatics analysis of the yeast phosphoproteome.

Since there is a substantial number of homologous proteins that is shared because of the evolutionary conservation between the yeast phosphoproteome and phosphoproteome of 'higher' eukaryotes such as fly, mouse and human (Gnad, Ren et al. 2007), observations from yeast phosphoproteome may help draw general questions about phosphorylation and understand certain biological processes (such as signal transduction) in other eukaryotes.

Comparative analyses of plant phosphoproteomes have shown that more than 50% of the phosphoproteins identified in rice and *Arabidopsis* had an orthologous phosphoprotein in the other species (Kersten, Agrawal et al. 2009; Nakagami, Sugiyama et al. 2010). Therefore, such a characterization of conserved phosphoproteome signatures as well as species-specific phosphorylation sites (non-conserved, adaptation mediated) will promote our understanding of core regulatory systems in plant species (Kersten, Agrawal et al. 2009; Nakagami, Sugiyama et al. 2010).

1.3.7 Post-translational regulation and genome evolution

As mentioned above, phosphorylation is perhaps the most common post-translational modification. Proteins with multiple phosphorylation sites can adopt several different functions, which largely depends on which site becomes occupied by phosphate and functional. When the conformation of a protein is changed by the phosphorylation of one of its particular amino acids, it can in turn allow different amino acids within the same protein being more accessible to kinases and thus promote more

phosphorylation events, or it can prevent the phosphorylation of nearby amino acids through sterical hindrance (Thingholm, Jensen et al. 2009). Therefore, it is of great interest to investigate how these phosphorylated sites evolved and why they were retained through millions of years of yeast genome evolution.

Gene and genome duplications create novel genetic material on which evolution can work and have therefore been recognized as major sources of innovation for many eukaryotic lineages. Whole genome duplications (WGDs), the simultaneous acquisition of extra copies of all the nuclear chromosomes of an organism, have occurred repeatedly in eukaryotic evolution and have been linked to genetic innovation, adaptation, speciation and survival (Freeling and Thomas 2006) (Ohno 1970; Freeling and Thomas 2006; Scannell, Byrne et al. 2006; Fawcett, Maere et al. 2009). WGDs, also referred as polyploidizations, have been reported to occur in all four eukaryotic kingdoms: plants (Simillion, Vandepoele et al. 2002; Blanc and Wolfe 2004; Yu, Wang et al. 2005; Tuskan, Difazio et al. 2006), animals (Amores, Force et al. 1998; McLysaght, Hokamp et al. 2002; Jaillon, Aury et al. 2004; Dehal and Boore 2005; Evans, Kelley et al. 2005) and fungi (Wolfe and Shields 1997; Dietrich, Voegeli et al. 2004; Kellis, Birren et al. 2004). An individual would usually posses two sets of genomes (called diploids), however, once in a while, due to errors in the reproduction process, individuals with four or more sets of genomes (corresponding to tetraploids or polyploids) will arise. However, polyploidy often has several disadvantages, or is even lethal in some cases (Comai 2005). Previous studies have uncovered multiple ancient polyploidization events in the evolutionary history of several angiosperm lineages. About 100 million years ago (mya), a WGD occurred in the hemiascomycetes yeast before the speciation of the budding yeast *Saccharomyces cerevisiae*.

Following a WGD, most duplicate copies are lost (Lynch and Conery 2000; Maere, De Bodt et al. 2005). For instance in yeast ~85% of copies have been lost since the WGD, but a considerable fraction survived, with either selection or genetic drift

accounting for the pattern of duplicate gene retention. However, the mechanisms that determine which genes are maintained in duplicate and which return to a single-copy state, are largely unknown. Therefore, I want to explore whether the rapid evolution of phosphorylation sites might be linked to gene retention. I have exploited the wealth of proteomic and genomic data available for the baker's yeast *Saccharomyces cerevisiae* and examined the relationship between protein phosphorylation, gene retention, and functional divergence following the WGD that occurred in the hemiascomycete yeasts, about 100 mya (Wolfe and Shields 1997; Kellis, Birren et al. 2004). This work will be presented in chapter 3 with great details.

Post-translationally modified amino acids (in particular phosphorylated serines) might have evolved differently from their non-modified counterparts. Previous studies have identified two phosphorylation-active forms: mutating Serine/Threonine (S/T) to Aspartic acid/Glutamic acid (D/E) has a similar effect to phosphorylation. In Chapter 4, using the yeast phosphoproteome, we want to study the pattern of pS substitutions through pairwise alignments between *S. cerevisiae* genes and their orthologs in other six fungal species.

2 Evaluation and properties of the budding yeast phosphoproteome

Amoutzias, G., He, Y., Lilley, K., Van de Peer, Y., Oliver, S.

Redrafted from a publication under revision in *Molecular Cellular Proteomics*, 2011

Abstract

We have assembled a reliable phosphoproteomic dataset for budding yeast *S. cerevisiae* and have investigated its properties. Twelve publicly available phosphoproteome datasets were triaged to obtain a subset of high-confidence phosphorylation sites (p-sites), free of 'noisy' phosphorylations. Analysis of this combined dataset suggests that the inventory of phosphoproteins in yeast is close to completion, but that these proteins may have many undiscovered p-sites. Proteins involved in budding and protein kinase activity have high numbers of p-sites and are highly over-represented in the vast majority of the yeast phosphoproteome datasets. The yeast phosphoproteome is characterised by a few proteins with many p-sites and many proteins with a few p-sites. We confirm a tendency for p-sites to cluster together and find evidence that kinases may phosphorylate off-target amino acids that are within 1 or 2 residues of their cognate target. This suggests that the precise position of the phosphorylated amino acid is not a stringent requirement for regulatory fidelity. Compared to non-phosphorylated proteins, phosphoproteins are more ancient, more abundant, have a higher probability of being essential, have longer unstructured regions, have more genetic interactions, more protein interactions and are under tighter post-translational regulation. It appears that phosphoproteins constitute the raw material for pathway rewiring and adaptation at various evolutionary rates.

2.1 Introduction

The application of mass spectrometry combined with affinity techniques that enrich for phosphopeptides has revolutionized the field of phosphoproteomics, such that hundreds, or even thousands, of phosphorylation sites (p-sites) may be identified in a single experiment. However, as with any high-throughput (HTP) technique, there are concerns about data quality and potential biases in the enrichment and identification procedures (Bodenmiller, Mueller et al. 2007; Lienhard 2008; Landry, Levy et al. 2009). Thus, there is a need for a stringent data evaluation to filter out possibly spurious p-sites before drawing any general conclusions about the structure and properties of a phosphoproteome.

There are several reasons for using yeast to benchmark these novel phosphoproteomics technologies. The most important of which is that a large number of phosphoproteomics experiments have been performed with *Saccharomyces cerevisiae*, under a reasonably wide range of conditions (Gruhler, Olsen et al. 2005; Chi, Huttenhower et al. 2007; Li, Gerber et al. 2007; Albuquerque, Smolka et al. 2008; Bodenmiller, Campbell et al. 2008; Beltrao, Trinidad et al. 2009; Gnad, de Godoy et al. 2009; Holt, Tuch et al. 2009; Huber, Bodenmiller et al. 2009; Soufi, Kelstrup et al. 2009; Stark, Su et al. 2010). Mass spectrometry-based proteomic methods sample the available proteome in a quasi-random manner (Aebersold and Mann 2003). Moreover, a large fraction (~80%) of the predicted yeast proteome has been found to be expressed under normal laboratory growth conditions, with high-throughput tagging or MS-based proteomics (Ghaemmaghami, Huh et al. 2003; de Godoy, Olsen et al. 2008; Wu, Dephoure et al. 2011). In a specific example that highlights the power of HTP proteomics, tandem MS approaches have managed to identify approximately 85-90% of all yeast mitochondrial proteins (Sickmann, Reinders et al. 2003). Finally, there is a wealth of relevant functional genomic information available for the organism, including data on protein abundance, half-lives, and the number of

kinases targeting a given protein (Ghaemmaghami, Huh et al. 2003; Ptacek, Devgan et al. 2005; Belle, Tanay et al. 2006; Newman, Ghaemmaghami et al. 2006), amongst others. All of these factors should assist in an in-depth bioinformatics analysis of the yeast phosphoproteome.

2.2 Experimental procedures

Chron. order	Author/date	Conditions	p-sites	p-proteins	Thresholds applied
1	Gruhler et al., 2005	alpha factor treated cells	676	470	Manually inspected by authors
2	Chi et al., 2007	?	724	422	P<1e-4
3	Li et al., 2007	Alpha factor arrested cells	1433	755	Ascores>=19; Peptide score=>45; dCn=>0.15
4	Albuquerque et al., 2008	DNA-damage response (MMS)	3155	1513	For Sequest: Xcorr=>1.8; p=<0.01; PLScore=>20. For Inspect: p=<0.01; PLscore=>20
5	Bodenmiller et al., 2008	?	2274	1071	Prophet>=0.99; dCn>= 0.15
6	Beltrao et al., 2009	Exponential growth in rich media	201	177	Peptide expect=<0.01; unambiguous psites only
7	Huber et al., 2009	Rapamycin/ Cycloheximide	311	160	Peptide probability=>0.99; dCn=>0.15
8	Holt et al., 2009	Asynchronous population	1939	857	Ascores>=19; dCn=>0.15
9	Holt et al., 2009	Arrested in mitosis with the spindle poison nocodazole	3348	1286	Ascores>=19; dCn=>0.15
10	Holt et al., 2009	Arrested in late mitosis by overexpression of a nondegradable cyclin, Clb2-delta	4321	1400	Ascores>=19; dCn=>0.15
11	Gnad et al., 2009	Grown for 10 generations on YNB (i.e. minimal) ?+glucose until they reached log-phase.	1546	726	Peptide probability=<0.01; Mascot=>30; psite probability=>0.99
12	Soufi et al., 2009	Osmotic stress	1155	682	Mascot=>30; psite probability=>0.99

Table 2.1. The twelve publicly available phosphoproteomic datasets of the 12HQ compendium.

Twelve high-throughput experiments between the years 2005-2009 were merged in the 12HQ dataset (where HQ stands for high quality). Details of these datasets may be found in Table 2.1. These experiments were performed with yeast cells in a variety of physiological and developmental states, including mating, exponential growth, different phases of the cell cycle and also challenged by DNA-damaging agents, osmotic stress, rapamycin, or cycloheximide. Apart from two cell-cycle experiments, asynchronous populations were analysed. There is substantial variation

between these datasets in terms of both yeast strains used and the analytical protocols employed to identify p-sites.

To ensure the high quality of the combined phosphoproteome dataset, we required that phosphopeptides were correctly identified with a probability of ≥99%, and that p-sites were correctly localized with a similar probability, in each experiment (see Table 2.1 for the thresholds applied to each public dataset). These are more stringent criteria than those used in the original published studies and, therefore, our 12HQ represents a high-confidence subset of the original data, which should also address the potential problem of false positive in some of the data (by false positives, we mean inaccurate assignment of p-sites by the MS identification technology).

2.3 Results & Discussion

Before analyzing the properties of the phosphoproteome, we needed to ensure that all experiments may be used and, therefore, a series of quality controls were performed.

2.3.1 No single dataset dominates the compendium

First, we determined whether analyses of 12HQ would be distorted by experiments with a relatively excessive number of p-sites dominating the combined dataset. By removing each of the original datasets individually from 12HQ, resulted in a 0-16% reduction in the number of p-sites and a 0-11% reduction in the number of phosphoproteins. Therefore, no individual experiment dominates 12HQ, and the degree of overlap between experiments provides further assurance of the quality of the combined dataset.

2.3.2 The various experiments significantly overlap with each other

Every published experiment identifies a number of p-sites that had been identified previously by other experiments. On average, it appears that, for any two experiments, ~12% of p-sites and ~28% of phosphoproteins are shared. The overlap observed between any two experiments is always statistically significant, whether it is for p-sites or the phosphoproteins identified (chi-square $p<0.05$). Therefore, there was no need to exclude any of these twelve experiments from our study. Interestingly, two experiments from different groups that were performed in very similar conditions (alpha-factor treated cells) (Gruhler, Olsen et al. 2005; Li, Gerber et al. 2007) had a much lower overlap (11% of p-sites & 31% of phosphoproteins) between them than two experiments of the same group that were performed in two different phases of the cell cycle (28% & 54% respectively) (Holt, Tuch et al. 2009). In terms of p-site identification, the implication is that the protocols used seem to be more important than the experimental conditions.

2.3.3 Saturation of the current phosphorylation dataset compendium for yeast

Next, we investigated the likelihood that 12HQ contains the majority p-sites and phosphoproteins that make up the entire yeast phosphoproteome. Figure 2.1 shows the incremental increase in the total number of non-redundant p-sites and phosphoproteins identified, following each HTP experiment. It is evident that the current compendium has not reached saturation in terms of p-sites. Moreover, the manually curated dataset from PhosphoGrid (Stark, Su et al. 2010), that contains only high-confidence functional p-sites from low-throughput (LTP) studies, supports

this conclusion. Only 27% (131/480) of the PhosphoGrid p-sites have also been identified in any of the HTP phosphoproteomics experiments. Stark et al. (Stark, Su et al. 2010) reached the same conclusions when they compared the PhosphoGrid dataset with a more limited yeast phosphoproteomics compendium. Huber et al. (Huber, Bodenmiller et al. 2009) reported that many rapamycin-sensitive phosphorylation events that were known from the literature could not be found in their HTP experiment. Finally, Albuquerque et al. (Albuquerque, Smolka et al. 2008) reported that the extent of phosphorylation of two low-abundance proteins, Rad9p and Mrc1p was 50% less in the HTP experiment compared to another experiment in which these two proteins were purified.

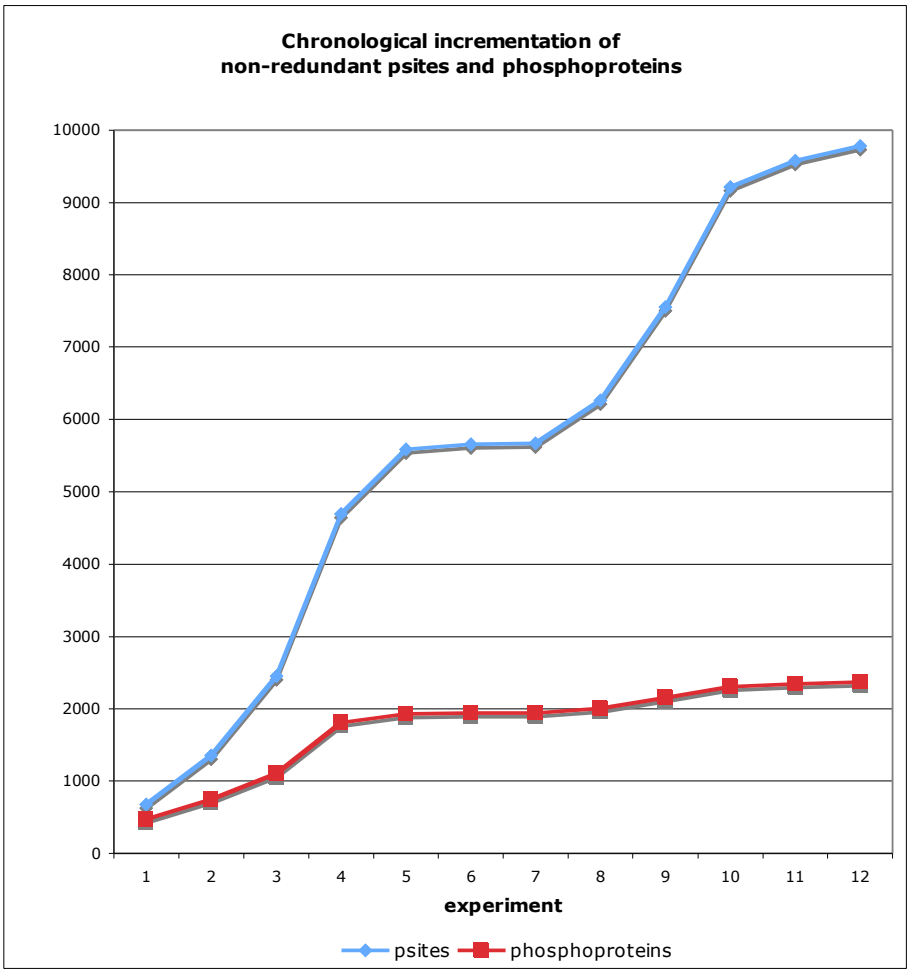


Figure 2.1. Incremental increase, with time, of the compendium for non-redundant p-sites and phosphoproteins.

Regarding the identification of phosphoproteins, Figure 2.1 suggests that the recorded phosphoproteome is slowly approaching saturation. Again, the PhosphoGrid dataset supports this conclusion, since 85% (122/144) of the LTP-identified phosphoproteins in PhosphoGrid were also identified by the HTP phosphoproteomics experiments. Furthermore, in a study that compiled fewer experiments, Beltrao et al (Beltrao, Trinidad et al. 2009) estimated that the detected yeast phosphoproteome from HTP studies is at the level of 81-92% saturation. Most probably, several phosphoproteins have yet to be detected, either because of their very low expression levels or because an insufficient number of biological conditions have been studied. Nevertheless, all the above facts converge on the notion that the majority of phosphoproteins has already been detected. The two trends in Figure 2.1 are further supported by the fact that the average overlap of p-sites, observed between any two experiments, is 12% compared to one of 28% for phosphoproteins.

The lack of saturation in terms of p-sites could be attributed to an insufficient number of environmental conditions having been tested in the experiments, or due to biases and weaknesses of the current phosphoproteomics technologies and protocols employed. Indeed, 57% of p-sites in the compendium have been identified only once. Interestingly, the majority of detected p-sites in a given experiment do not seem to be regulated in that specific condition. Gruhler et al. (Gruhler, Olsen et al. 2005) reported that only 18% of the detected phosphopeptides were regulated by alpha factor in their experiment, whereas Soufi et al. (Soufi, Kelstrup et al. 2009) reported that 15% of detected p-sites changed status after osmotic shock treatment. Similar conclusions were reached by Huber et al. (Huber, Bodenmiller et al. 2009) in another experiment with rapamycin treatment. In addition, Huber et al. observed that rapamycin-sensitive p-sites that had been rigorously defined by LTP experiments were not detected in their HTP experiment. This is a clear demonstration that these technologies and protocols require further development and refinement.

2.3.4 The non-phosphoproteome

We wanted to investigate if there are any basic differences between phosphorylated and non-phosphorylated proteins that may affect the detection of p-sites. It is important to determine if any underlying differences have a biological basis or whether they stem from biases in the MS technologies or other experimental protocols used. Therefore, we identified a collection of yeast proteins for which there is no extant evidence of their being phosphorylated in any of the 12 HTP experiments, even if we do not apply any filters at all. We call this collection of proteins the non-phosphoproteome; it is composed of 2219 ORFs.

It is conceivable that the non-phosphoproteome is an artefactual dataset that merely contains proteins that are inherently undetectable by high-throughput (HTP) proteomic Mass-Spectrometry (MS) technologies. In order to account for this potential inherent undetectability, we also generated a subset (1418 out of the 2219 proteins) of the original non-phosphoproteome that was actually detectable by HTP-MS proteomics (designated as MS-detectable non-phosphoproteome). To this end, we used two HTP yeast proteomic datasets that were detectable by MS technology (de Godoy, Olsen et al. 2008; Wu, Dephoure et al. 2011). These two HTP-MS experiments, when combined together, identified 4656 yeast proteins in total, where 86% of them are found in both datasets. This is a strong confirmation of the reproducibility of the MS technology for protein detection, even by different laboratories. In this study, any analyses performed with the non-phosphoproteome were also performed with the MS-detectable non-phosphoproteome, to control for protein detectability.

GO-slim analysis with Bingo (Maere, Heymans et al. 2005) on the non-phosphoproteome revealed a statistically significant enrichment in proteins found in the mitochondria, membranes, cell wall, endoplasmic reticulum, and the extracellular space (the above conclusions, with the exception of those for the cell wall, are also supported by the MS-detectable non-phosphoproteome). Nevertheless, membrane

proteins are considered more difficult to detect by mass spectrometry, than are cytosolic proteins. Gnad et al. (Gnad, de Godoy et al. 2009) also reported that their dataset was underrepresented in mitochondrial and endoplasmic reticulum proteins. We investigated a small dataset from Reinders et al. (Reinders, Wagner et al. 2007) that was specifically designed to detect phosphorylation events in the mitochondrial fraction of the proteome. Previous analyses have detected ~850 proteins in mitochondria, with a coverage of 85% of known mitochondrial proteins (Sickmann, Reinders et al. 2003; Reinders, Zahedi et al. 2006; Reinders, Wagner et al. 2007). From the 78 p-sites found in 46 proteins, 22 p-sites (28%) and 24 proteins (52%) were also detected in the 12HQ dataset. Therefore, we believe that the under-representation of mitochondrial proteins in the HTP phosphoproteome generated by MS analyses has a biological basis. This finding may relate to the prokaryotic origin of mitochondria and the recent observations (Macek, Mijakovic et al. 2007; Macek, Gnad et al. 2008; Soufi, Gnad et al. 2008; Lin, Hsu et al. 2009; Ravichandran, Sugiyama et al. 2009; Sun, Ge et al. 2009; Parker, Jones et al. 2010; Prisic, Dankwa et al. 2010; Schmidl, Gronau et al. 2010) that prokaryotic proteins are not as extensively phosphorylated as are those of eukaryotes.

2.3.5 The impact of the abundance and half-life of proteins

Next, we investigated if protein abundance is a confounding factor for the detection of a phosphoprotein in the MS experiments and, for this purpose, we used three comprehensive protein abundance datasets (Ghaemmaghami, Huh et al. 2003; Newman, Ghaemmaghami et al. 2006). We compared protein abundances of the phosphoproteome (2781 proteins) against the non-phosphoproteome (2219 proteins) and observed that the phosphoproteome had, on average, a 2-3 times higher abundance than the non-phosphoproteome dataset. This result was consistent for each of the three abundance datasets (Wilcoxon $p<3e-7$). We also determined whether, in each of the 12 experiments, the detected phosphoproteins were members of the higher abundance classes. In 78% of the cases (3 abundance

datasets for 12 MS experiments), we found the phosphoproteins to have significantly higher abundance (Wilcoxon $p<0.05$). Beltrao et al. (Beltrao, Trinidad et al. 2009) also reported that phosphopeptides were 3 times more abundant than non-phosphorylated proteins, but regard this as a small difference, given the 8 orders of magnitude span in protein abundances (Beltrao, Trinidad et al. 2009). Although a bias clearly exists, we also believe that it is not the determining factor for identifying a phosphoprotein because there is no clear distinction in the level of abundance of the phosphoproteome versus the non-phosphoproteome. The abundances of both groups of proteins span a similar range of orders of magnitude (see Supplementary Information). The above conclusions are also supported when accounting for HTP MS-detectability (see Supporting Information).

Another factor that might affect the detection of a phosphoprotein is its half-life, because rapid degradation could make a phosphoprotein more difficult to detect. The data indicate, while protein turnover is a consideration, it is not a major one. When we analysed a comprehensive yeast protein half-life dataset (Belle, Tanay et al. 2006), we found that the 12HQ phosphoproteins had, on average, a 50% lower half-life than non-phosphoproteins (Wilcoxon $p<0.0011$). The above conclusions are also supported when accounting for HTP MS-detectability (see Supporting Information, Controlling for MS-detectability).

2.3.6 The importance of protein structure

Protein kinases have a high preference for phosphorylating serine, threonine & tyrosine (STY) amino acids that are embedded within intrinsically disordered regions (Iakoucheva, Radivojac et al. 2004). We wanted to investigate whether non-phosphorylated proteins lacked disordered regions and, for this purpose, the intrinsic disorder (ID) of yeast proteins was predicted (Peng, Radivojac et al. 2006). The 12HQ phosphoproteins had, on average, ID regions that were 182% longer than

those of the non-phosphoproteins (an average ID length of 330 and 117 residues per protein for phosphoproteins and non-phosphoproteins respectively; this is statistically significant - Wilcoxon p=0). Interestingly, the 12HQ phosphoproteins had, on average, non-ID regions that were 38% longer than those of the non-phosphoproteins (an average non-ID length of 294 and 214 residues per protein for phosphoproteins and non-phosphoproteins, respectively; this is statistically significant - Wilcoxon p<2e-16). The above conclusions are also supported when accounting for HTP MS-detectability (see Supporting Information, Controlling for MS-detectability). In each of the 12 experiments, proteins that were detected as phosphorylated (even if they were excluded from the 12HQ dataset) always had longer ID regions than non-detected (i.e. non-phosphorylated) proteins (Wilcoxon p=0). Thus a strong bias exists, and there is a substantial difference in the total length of disordered regions in phosphoproteins compared to non-phosphoproteins (see Figure 2.2). In addition, we observed a moderate correlation (Pearson coefficient = 0.55) between the number of p-sites on a protein and the length of its ID region. It should be noted that ID regions are not only involved in interactions with kinases, but in transient protein interactions in general (Gsponer, Futschik et al. 2008).

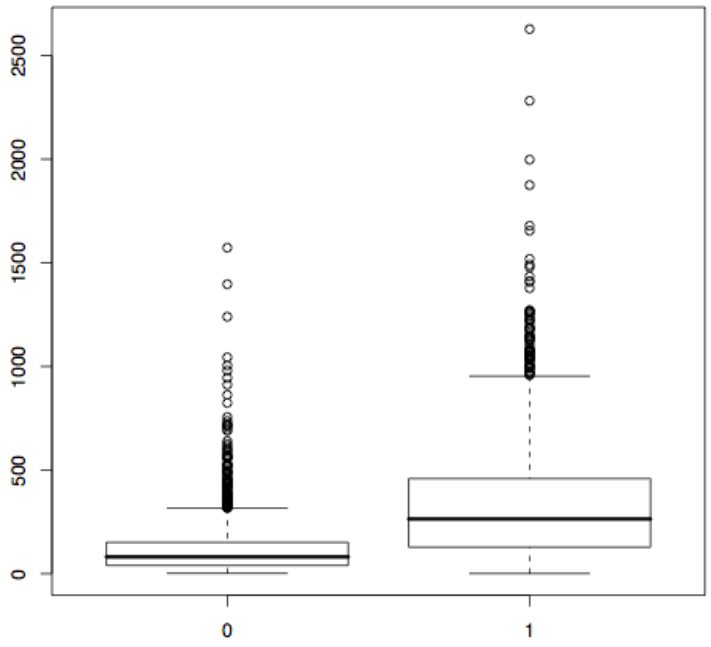


Figure 2.2. Boxplot of the length (in amino acids) of regions of intrinsic disorder, for the non-phosphoproteome (denoted with 0) and the 12HQ phosphoproteome (denoted with 1).

2.3.7 Peptide analysis

We next investigated whether there were likely to be differences in the length or relative charge of the digested peptides generated for MS analysis between phosphorylated and non-phosphorylated proteins. It is conceivable that the current protocols can detect only a narrow spectrum of the phosphopeptides that is not present in the negative dataset. We thus performed a theoretical trypsin digestion of the proteins in the two datasets with the proteogest tool (Cagney, Amiri et al. 2003) and calculated the length and relative charge of the theoretical peptides. We did not observe any substantial difference in the distribution of either variable between the two datasets (see Supporting Information). The above conclusions are also supported when accounting for HTP MS-detectability (see Supporting Information, Controlling for MS-detectability).

2.3.8 Functionality of p-sites and biological noise

Recently, concerns have been raised about the functionality of p-sites detected in analyses using MS and that the importance of biological noise has been underestimated in these HTP experiments (Lienhard 2008; Landry, Levy et al. 2009). Lienhard has raised the possibility that, due to the high sensitivity of these MS instruments, biologically noisy p-sites are being detected (Lienhard 2008). ‘Biological noise’, in this case, represents phosphorylation events occurring in degenerate motifs by non-cognate kinases; frequent (but low abundance) off-target phosphorylations, etc. Landry et al. exploited evolutionary information to estimate that up to 65% of p-sites in these HTP experiments could be non-functional, thus indicating that biological noise could be a significant problem (Landry, Levy et al. 2009).

The presence of such a high number of HTP MS experiments for yeast allows us to address this very important issue. First of all, a basic assumption is made, that a p-site identified in many experiments is probably not due to stochastic off-target kinase interactions but, rather, has a high probability of being functional. Several factors could possibly invalidate this basic assumption, such as the inherent detectability of certain proteins or p-sites, driven by protein abundance, modification stoichiometry, peptide properties etc. In addition, we cannot exclude the possibility that the functional effect of particular phosphorylation events is entirely neutral. Nevertheless, strong indications of the validity of this basic assumption come from five independent analyses, shown below:

First of all, the 12HQ compendium was compared to a list of proteins that might be phosphorylated in any of the 12 MS experiments but did not meet our stringent filtering criteria; the 12HQ set was found to be more enriched in PhosphoGrid proteins (5.1% vs 1.7% respectively; chi-squared p=5e-7). Within the 12HQ compendium, we compared a list of proteins identified as phosphorylated in 3 or more experiments vs another list of proteins identified as phosphorylated in 1 or 2 experiments and found that the first list was more enriched in PhosphoGrid proteins (6.4% vs 3.4% respectively; chi-squared p=0.0015). Therefore, as our filtering criteria become more stringent, the corresponding datasets are also becoming more enriched in proteins from the ‘gold-standard’ PhosphoGrid dataset, which is compiled from low-throughput experiments and so is not affected by biases of the high-throughput experiments.

Second, we wanted to exclude the possibility that protein abundance, or the length or chemistry of a digested peptide, affected the number of times a protein was detected as phosphorylated in our 12HQ compendium. To examine this possibility, we binned proteins in 12 groups, depending on how many times they were found to be phosphorylated. These bins were compared for each of the above three properties

(abundance, peptide length, peptide relative charge), but no significant differences were found between the different bins (see Supporting Information).

Third, we investigated whether p-sites identified in only a few experiments tend to be found within more degenerate motifs than p-sites identified in many experiments. For this analysis, we used the netphosyeast prediction algorithm that is considered the best performing algorithm for yeast motifs (Ingrell, Miller et al. 2007). Netphosyeast makes predictions for serines and threonines, but not for tyrosines. As a negative comparator, we used the 195,109 ST amino acids in the 12HQ proteins, for which there is no evidence of phosphorylation, even if we do not apply any filters on the data. This collection of ST amino acids constitutes the no-p-sites dataset, symbolised with zero in Figure 2.3. As a positive comparator, we also used the 473 ST amino acids in the PhosphoGrid dataset; these are known to be phosphorylated and functional. This collection of ST amino acids is symbolised with 13 in Figure 2.3. It is evident that p-sites with a coverage of 2x or more have very similar median prediction scores to that of the PhosphoGrid set of known functional p-sites. Furthermore, the differences in netphosyeast scores were statistically different among the various adjacent bins of p-site coverage (negative vs 1x, 1x vs 2x; Wilcoxon test $p<2e-16$ & $p<2e-16$ respectively).

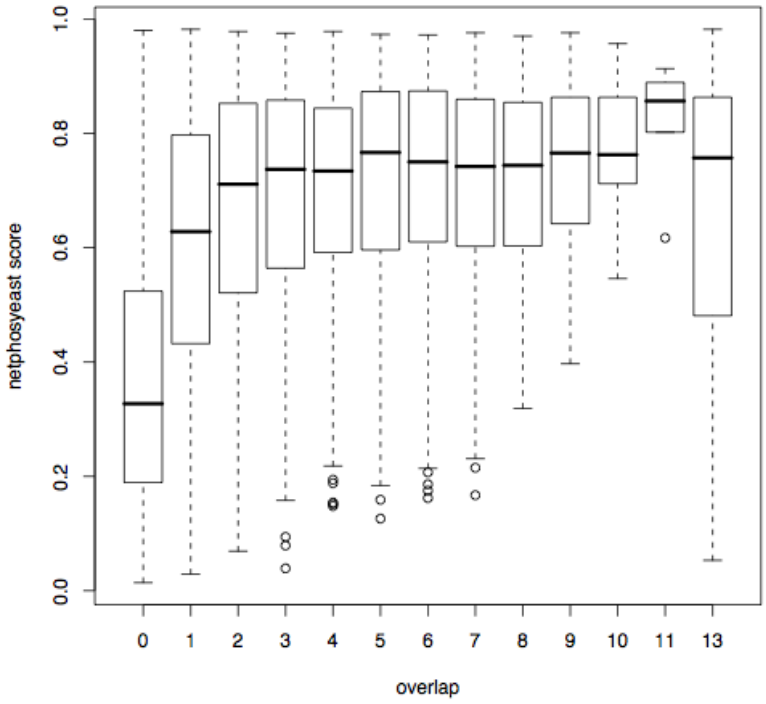


Figure 2.3. Boxplot of netphosyeast prediction scores for p-sites with a certain coverage (number of times identified). **0** = the 195,109 ST amino acids in the 12HQ proteins, for which there is no evidence that they are phosphorylated, even when no filters are applied. **1-12** = the number of experiments in the 12HQ set in which an ST amino acid has been detected as phosphorylated. **13** = the 473 ST amino acids known to be phosphorylated and functional according to the PhosphoGrid dataset.

Fourth, the same conclusions about motif degeneracy are reached when we analyse the predictions supplied by Mok et al. (Mok, Kim et al. 2010). This group used a peptide library approach to determine consensus phosphorylation site motifs for almost half the yeast protein kinases (61/122). By integrating these sequence motifs together with other features, such as evolutionary conservation, disorder and protein surface accessibility, they used a Bayesian algorithm (MOTIPS) to predict phosphorylation sites for certain kinases. By applying a likelihood threshold of >0.5, we observed that MOTIPS could assign a kinase to 34% (165/480) p-sites in PhosphoGrid, 40% (1027/2566) p-sites in 12HQ_3x (p-sites that have been detected in 3 or more experiments), 34% (3359/9783) p-sites of 12HQ, and 15% (34916/239269) of the non-p-sites in the 12HQ dataset. Thus the Mok et al. predictions, which are independent from those of netphosyeast, confirm that p-sites with higher coverage have both higher prediction scores from netphosyeast and more predicted phosphorylation motifs from MOTIPS.

Fifth, an additional indication that our assumption holds is the observation that, for disordered regions, p-sites detected in 3 or more experiments evolve 8% more slowly than p-sites detected in only 1 experiment (Wilcoxon p<9e-6). In addition, for disordered regions, p-sites detected in 3 or more experiments evolve 10% more slowly than non-phosphorylated S/T/Ys from 12HQ proteins (Wilcoxon p<3e-12). For this analysis, we used the evolutionary rates calculated by Landry et al. (Landry, Levy et al. 2009).

An important question, then, is: in how many experiments should a p-site have been discovered in order to confidently designate it as functional? To address this, we simulated the 12 phosphoproteomic experiments by shuffling the positions of the p-sites. For the simulation, we took into account the structure of the proteins (order/disorder), the number of STY amino acids in each protein, and the number of phosphorylation events detected in each experiment. 1000 simulations were performed and the results with the highest (by chance) coverage were retained for comparison with the observed coverage in the real dataset. Figure 2.4 shows the cumulative distribution of the coverage of p-sites for both the 12 experiments and for the simulation. In essence, Figure 2.4 investigates how many repeated observations of low-stoichiometry off-target p-sites we would expect to find by chance, if all of the phosphorylations were low-stoichiometry off-target events.

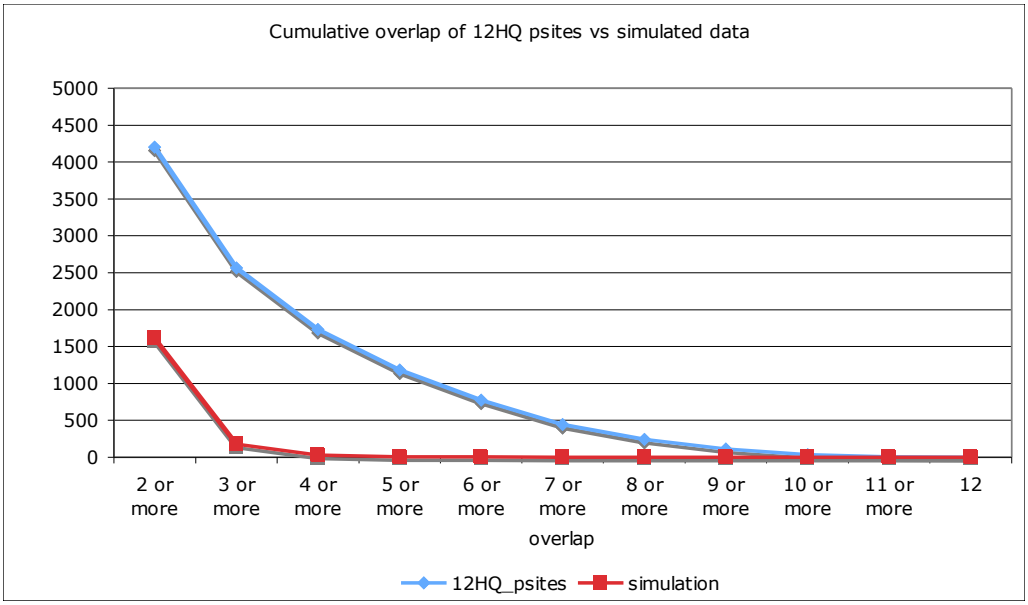


Figure 2.4. Cumulative distribution of coverage for p-sites of the 12HQ dataset and for the simulation.

We observed that, if all phosphogroups were assigned in a totally random manner, then we would expect 1614 p-sites with a coverage 2x or more, just by chance, whereas the observed number is 4204. The equivalent percentage for ≥3x coverage would be 177 expected and 2566 observed, whereas for ≥4x coverage it would be 27 expected and 1734 observed. Therefore, it seems reasonable to select ≥3x coverage as a very stringent cut-off in order to filter out off-target phosphorylation events. This cut-off does not necessarily mean that any specific p-site that has been identified only once or twice represents an example of low-stoichiometry off-target phosphorylation event. It only provides a very conservative and confident subset, that we designate 12HQ_3x.

The distribution in Figure 2.4 reveals that a substantial number of p-sites are found in many experiments. According to Soufi et al. (Soufi, Kelstrup et al. 2009) this could be explained by the asynchronous state of the cell populations in most of the experiments. However, the relatively high overlap (28% and 54% for p-sites and phosphoproteins respectively) in the 2 Holt et al. experiments (Holt, Tuch et al. 2009), which characterised the phosphoproteome at two different stages of the cell cycle,

indicates that this cannot be a complete explanation. We would suggest that some p-sites are ubiquitously in an ‘ON’ state (phosphorylated). It may be that the cell keeps a small percentage of the expressed protein molecules of a gene in this phosphorylated state and that this percentage changes according to external stimuli.

Having excluded technical noise, and having validated the high quality of the 12HQ dataset, we were in a position to investigate the properties of the yeast phosphoproteome.

2.3.9 General characteristics of the phosphoproteome

The compilation of the above 12 experiments leads to the 12HQ dataset, with 9783 p-sites found in 2374 phosphoproteins and the 12HQ_3x dataset with 2566 p-sites in 1112 phosphoproteins. Recently, Yachie et al. (Yachie, Saito et al. 2011) analysed a compendium of ~3,500 phosphoproteins containing 26,000 p-sites. We accredit this discrepancy in the absolute numbers of p-sites and phosphoproteins mainly to the very rigorous protocol we applied to filter out technical false-positive and low-abundance off-target phosphorylations, which are a major concern (Lienhard 2008; Landry, Levy et al. 2009). 17% of 12HQ p-sites and 12% of 12HQ_3x p-sites are found inside or in the vicinity (+/- 10 amino acids) of an annotated Pfam domain (we excluded Pfam-B domains, which are un-annotated and automatically generated). Serines, threonines and tyrosines constitute, respectively, 81%, 17% and 2% of the phosphorylated residues in yeast. This pattern of site preference is consistent across all of the 12 HTP experiments as well as the PhosphoGrid dataset, albeit with some variation. Therefore, we do not consider it an artifact of the MS technologies. The very low percentage of phosphorylated tyrosines is explained by the lack of tyrosine kinases in yeast and the dual specificity of certain kinases that may phosphorylate some tyrosines (Manning, Plowman et al. 2002; Li, Gerber et al. 2007).

Previous analyses show that most p-sites are found in disordered regions (Landry, Levy et al. 2009). Indeed, we also observed that 91% of STY p-sites are found in disordered regions, compared to 54% of non-phosphorylated STY sites. In PhosphoGrid, the percentage is very similar, around 92%. The structural properties of the region around the p-site probably play an important functional role. It is considered that kinases tend to phosphorylate sites that are easy to access, thus it makes sense that p-sites should be embedded within unstructured regions. These conclusions are robust for the 12HQ_3x subset as well.

2.3.10 Functional analysis using GO-slim

Functional categories that are related to cell budding and kinase activity are highly over-represented in the vast majority of the HTP experiments. Furthermore, a GO-Slim enrichment heat-map (see Supporting Information) reveals that the functional categories of proteins that are usually found to be phosphorylated are very similar among the various experiments. Our findings are in agreement with those of Beltrao et al. (Beltrao, Trinidad et al. 2009) using data for three different yeast species (*S. cerevisiae*, *Candida albicans*, *Schizosaccharomyces pombe*). Their analysis showed functional categories such as budding, cytokinesis, and signal transduction to be over-represented in the phosphoproteins of all three species. This consistency in terms of GO-Slim categories is in contrast to the statistically significant, but nevertheless rather moderate, overlap of p-sites between the 12 S. cerevisiae HTP experiments. Apparently, in every HTP experiment, different p-sites are found to be phosphorylated, but usually either on the same proteins or on proteins within the same functional categories. This conclusion may be biologically meaningful, but artefacts due to the range of physiological and developmental conditions studied, or the experimental protocols employed to identify phosphorylated proteins, cannot be excluded.

2.3.11 Distribution of p-sites in yeast proteins

On average, we found 4 p-sites per phosphoprotein in the 12HQ dataset. Proteins with functions involved in cell budding, cytoskeleton, and signal transduction have a higher than average number (6-8) of p-sites per phosphoprotein. The distribution of p-sites in phosphoproteins is markedly skewed (Figure 2.5). Most phosphoproteins have a small number of p-sites, whereas a very small number of phosphoproteins have many p-sites. For example, the top 10 most phosphorylated proteins have between 32-54 p-sites each. When we use the more stringent 12HQ_3x dataset, we observe the same skewed distribution, but the top 10 most phosphorylated proteins have between 10-23 p-sites. Due to the incompleteness of the phosphoproteomics datasets we expect the actual number of p-sites to be significantly higher. The most phosphorylated protein (with 54 p-sites) is Sec16p (YPL085W), which is a coat protein of the COPII vesicle, required for ER transport (Supek, Madden et al. 2002). A similar distribution has been reported for other species, such as human and mouse (Yachie, Saito et al. 2009).

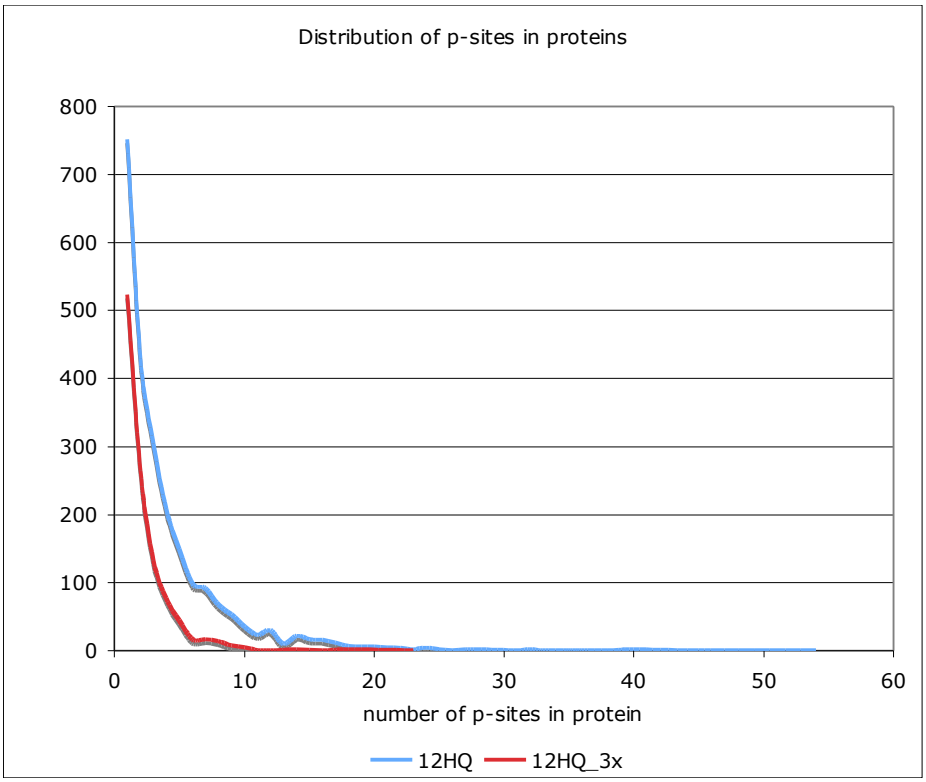


Figure 2.5. Distribution of p-sites in proteins, for the 12HQ and 12HQ_3x datasets.

2.3.12 Phosphoproteins are of more ancient origin than non-phosphorylated proteins

In a previous analysis with a smaller dataset, Chi et al. (Chi, Huttenhower et al. 2007) observed that phosphoproteins tend to be of more ancient evolutionary origin than randomly chosen proteins. By using a published dataset of yeast orthologous groups (Wapinski, Pfeffer et al. 2007) that was based on phylogenetic analysis and chromosomal synteny from YGOB (Byrne and Wolfe 2005), we identified orthologs in the genomes of nine other fungi and observed that a higher fraction of *S.cerevisiae* phosphoproteins have orthologs in other fungal genomes than do non-phosphorylated proteins (see Figure 2.6). The above conclusions are also supported when analysing the 12HQ_3x dataset and also when accounting for HTP MS-detectability (see Supporting Material, Controlling for MS-detectability).

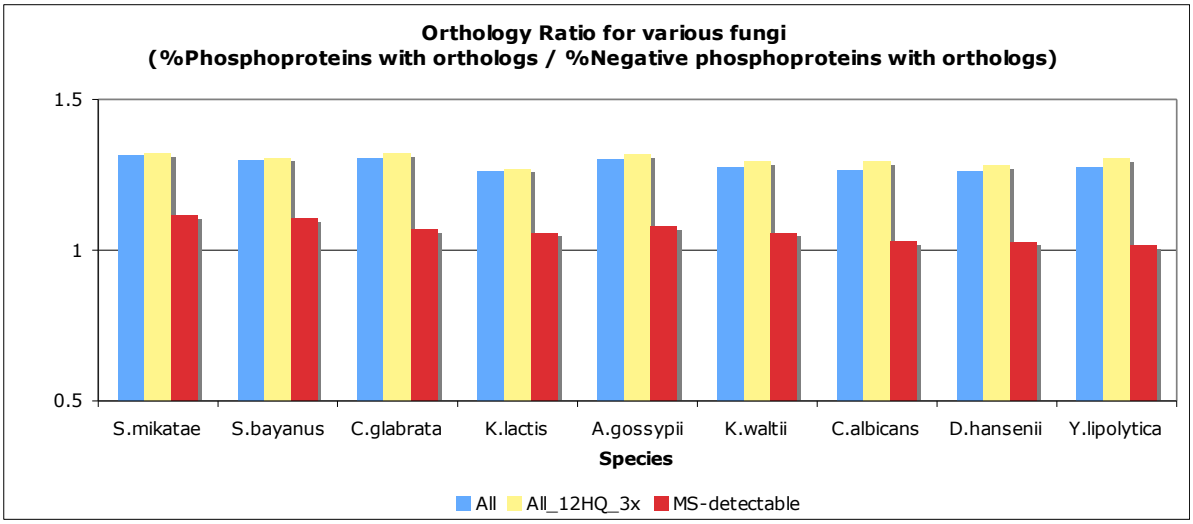


Figure 2.6. Orthology ratio for various fungi (orthologs based on phylogeny and YGOB) for the 12HQ vs negative dataset (blue color), for the 12HQ_3x vs negative dataset (yellow color) and the MS-detectable 12HQ vs negative dataset (red color). A ratio value >1 indicates that phosphoproteins have more yeast orthologs than the negative dataset, in that particular fungus.

2.3.13 Phosphoproteins are more frequently essential for yeast cell growth than non-phosphorylated proteins.

For this analysis, we used a high-confidence list of 956 well-defined essential genes (Giaever, Chu et al. 2002; Steinmetz, Scharfe et al. 2002; Pache, Babu et al. 2009). We observed that 23% of phosphoproteins are products of these essential genes, compared to 10% of non-phosphoproteins (Chi-squared < 6e-23). Nevertheless, the relationship between phosphorylation and essentiality is not a simple one, because only 17% of proteins with many p-sites (≥10) are essential. Thus a high number of phosphorylation sites on a protein does not necessarily increase the likelihood of its being essential. Indeed, in another study (see Chapter 3), we have shown that genes encoding phosphoproteins are more likely to be maintained in duplicate following whole-genome duplication events (Amoutzias, He et al. 2010). Our findings seem to contradict a previous analysis by Chi et al. (Chi, Huttenhower et al. 2007), done with a much smaller dataset, where there was no difference in essentiality between phosphoproteins and randomly selected proteins. The above conclusions are also

supported when accounting for HTP MS-detectability (see Supporting Material, Controlling for MS-detectability).

2.3.14 Phosphoproteins are under tighter regulatory control than non-phosphorylated proteins

Gsponer et al. (Gsponer, Futschik et al. 2008) have independently shown that unstructured proteins are under tight regulation at many levels of gene expression; therefore, we wanted to investigate if phosphoproteins are under tighter regulatory control than non-phosphorylated proteins. We analysed functional data such as the number of TFs that bind to the promoters of genes (Lee, Rinaldi et al. 2002; Harbison, Gordon et al. 2004; Balaji, Iyer et al. 2008), protein half-lives (Belle, Tanay et al. 2006), protein ubiquitination (Peng, Schwartz et al. 2003), high-confidence genetic (Costanzo, Baryshnikova et al. 2010) or protein-protein (Batada, Reguly et al. 2006) interactions and the number of different kinases targeting a protein (Ptacek, Devgan et al. 2005). We observed no statistically significant difference in the number of TFs that bind the promoters of genes that encode phosphoproteins compared to those that encode non-phosphorylated proteins. Nevertheless, phosphoproteins have, on average, 43-50% shorter protein half-lives (Wilcoxon $p<0.0011$) than non-phosphorylated proteins. In addition, a higher fraction of phosphoproteins are ubiquitinated, compared to the non-phosphoproteome (27% and 9% respectively; Chi-square $p<2e-16$). Phosphoprotein genes have, on average, 39-40% more genetic interactions than the non-phosphorylated dataset (Wilcoxon $p<1.6$ e-15). Furthermore, phosphoproteins have 45-48% more protein-protein interactions (Wilcoxon $p<2e-13$) than non-phosphoproteins, in accordance with a recent analysis on another yeast dataset (Yachie, Saito et al. 2011). Additionally, we observed a moderate correlation between the number of p-sites on a protein and the number of proteins interacting with it (Spearman coefficient = 0.3, for the 12HQ dataset). All of the above conclusions hold for the 12HQ_3x dataset and even when accounting for HTP MS-detectability (see Supporting Information). Recently, Shou et al. (Shou,

Bhardwaj et al. 2011), demonstrated that different types of molecular networks rewire at different rates with their order being (from fast to slow) transcriptional, phosphorylation, genetic interaction, miRNA, protein interaction and metabolic pathway networks. Apparently, phosphoproteins are enriched in those elements that act as the evolutionary raw material for adaption at various speeds.

2.3.15 Weak correlation between the number of phosphorylation sites on a protein and the number of different kinases that target it

For this analysis, we used the *in vitro* protein array experiment of (Ptacek, Devgan et al. 2005). As expected, the phosphoproteins dataset (2374 proteins) had over 3 times more kinase interactions than the non-phosphoprotein dataset (2219 proteins); Wilcoxon p=0. Interestingly, the number of kinases interacting with a phosphoprotein did not correlate strongly with the number of p-sites found in the protein (Pearson coefficient = 0.18). One potential explanation is that the specific experiment that measures which kinases target a protein is noisy because it is performed *in vitro*. Nevertheless, there exists a statistically significant overlap among phosphoproteins found by MS experiments and by the *in vitro* protein array experiments (687 proteins found in both datasets phosphorylated; chi-squared <2.2e-16). Further support for the *in vitro* approach comes from the fact that, for each of the 12 experiments, the proteins identified as phosphorylated were found to be targeted by 2.2-2.9 times more kinases than the proteins that were not identified as phosphorylated (Wilcoxon p<7e-10) by the HTP *in vivo* approach. A second, and more plausible, explanation is that there is no 1:1 relationship between kinases and phosphorylation sites. One kinase may phosphorylate many p-sites in a protein, or the same p-site may be phosphorylated by several closely related kinases. Indeed, Schweiger and Linial showed that groups of neighbouring p-sites in a protein may be phosphorylated by the same kinase (Schweiger and Linial 2010).

2.3.16 Clusters of p-sites

For more than half (51%) of 12HQ p-sites, there exists another p-site within a distance of ≤8 amino acids; for more than a third (33.5%) of the 12HQ p-sites, the interval is even smaller, ≤3 amino acids. Schweiger and Linial (Schweiger and Linial 2010) have demonstrated, for a different dataset, that this clustering is statistically significant. We repeated their analysis with the 12HQ and 12HQ_3x high quality datasets to ensure that this clustering is not an artefact resulting from the MS analysis mistakenly assigning the phosphogroup to a neighbouring STY amino acid. Such mistakes would be expected to be more common for the neighbouring amino acids of the most frequently phosphorylated p-sites. For our analysis, we used p-sites with a very high probability of correct localization (>99% for each p-site), therefore, our dataset tackles this very important issue.

Simulations were performed where we shuffled the p-sites in a protein, but did not change either the overall amino acid sequence, the distribution of p-sites found between disordered and ordered regions, or the total number of p-sites on the protein. 1000 simulations with the 12HQ dataset showed clearly that p-sites tend to cluster together more frequently than would be expected by chance (Figure 2.7), as Linial & Schweiger originally reported (Schweiger and Linial 2010). The simulations were repeated for the 12HQ_3x dataset, with the same conclusions.

Figure 2.7. Cumulative distribution of distance among neighbouring p-sites.

It has been suggested that the clustering of p-sites allows a greater flexibility of control over a protein's activity, permitting variability in the sensitivity or rapidity of the response to different intra- or extra-cellular stimuli (Nash, Tang et al. 2001; Gunawardena 2005; Schweiger and Linial 2010). A more mundane, but perfectly feasible, alternative is that this clustering may partly result from misphosphorylations made by the protein kinase. It is known that protein kinases recognize very degenerate target sequences and that the specificity inside the cell is determined by many other factors (e.g. co-expression, co-localization, scaffolding, etc) (Linding, Jensen et al. 2007; Won, Garbarino et al. 2011). Thus, most of the time, a kinase would detect its cognate motif and phosphorylate the correct STY amino acid. However, if there were another STY amino acid close by, then the kinase might phosphorylate this second residue in error. The error will depend on the quality of the motif. The more degenerate the non-cognate motif, the less frequently an off-target phosphorylation will occur.

We measured the distance between all adjacent pairs of p-sites for the 12HQ dataset and, for each of these pairs, we determined the log2 ratio of netphosyeast score and the log2 ratio of 12HQ p-site coverage. We reasoned that, if some kinases phosphorylate a neighbouring (off-target) ST amino acid, then the more degenerate the motif (lower netphosyeast score), the lower the probability its being the subject of an off-target phosphorylation. We tested the correlation for various distances (e.g. 1 residue, 2 residues, etc…) against a background correlation of distance ≥20 residues. We also calculated the 95% confidence intervals for all these correlations and found that, for the 12HQ dataset and a distance of 2 residues, the correlation is quite high (Pearson coefficient = 0.48) and also significantly higher than the background correlation (Pearson coefficient = 0.22) (see Figure 2.8). We infer that some kinases may be prone to making such neighbouring off-target phosphorylations, especially for amino acids within 2 residues of the cognate site. Interestingly, Schweiger and Linial indicate that the most prevalent distances in the observed clustering were 1-4 amino acids. We believe that these inherent neighbouring off-target phosphorylations of kinases are tolerated by the cell because the precise positioning of phosphorylation sites is not always required for proper regulation (Moses, Liku et al. 2007), thus highlighting the robustness of this molecular network.

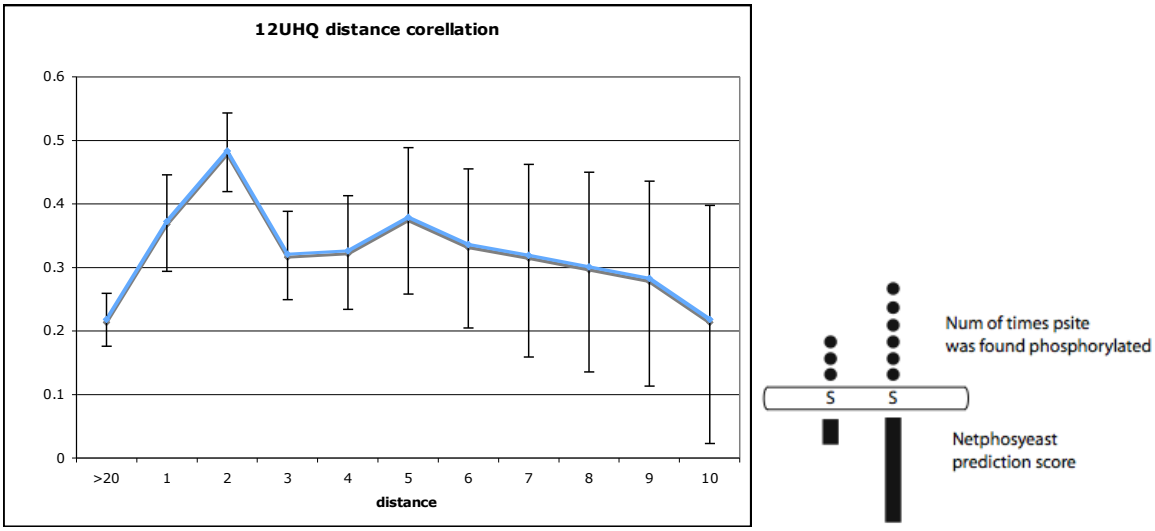


Figure 2.8. Plot of the Pearson coefficient values for $\log_2$ ratio of netphosyeast score and $\log_2$ ratio of coverage of p-sites. The correlation was calculated for neighbouring p-sites with a certain distance. For example, for neighbouring p-sites with a distance of 2 amino acids, we observe that the p-site with the higher netphosyeast score is also found in more experiments than the other p-site and that the correlation is quite high (0.48).

2.4 Conclusions

In this analysis, we integrated 12 HTP phosphoproteomic datasets from S. cerevisiae, published between the years 2005 – 2009, together with literature-curated LTP data from the PhosphoGrid database. We applied very stringent criteria to filter out both technical false-positives and low-stoichiometry off-target phosphorylations, which are a major concern (Lienhard 2008; Landry, Levy et al. 2009), and one that is not addressed properly in many analyses. We have thus provided a high quality dataset of p-sites, which may be employed to study the general properties of the yeast phosphoproteome. Our quality controls demonstrated that every HTP experiment correctly captured a fraction of the yeast phosphoproteome, but there is still plenty of room for improvements in the technologies and protocols used. The compendium may well be approaching saturation in terms of identifying all yeast phosphoproteins, but it is far from complete in terms of identifying all the p-sites on those proteins.

The yeast phosphoproteome is characterised by a few proteins with many p-sites and many proteins with a few p-sites. Proteins involved especially in budding and protein kinase activity have high numbers of p-sites. We confirm a tendency for p-sites to cluster together and find evidence that kinases may be involved in low-stoichiometry off-target phosphorylations of amino acids that are within 1 or 2 residues of their cognate target. This suggests that the precise position of the phosphorylated amino acid is not a stringent requirement for regulatory fidelity. Compared to non-phosphorylated proteins, phosphoproteins are more ancient in evolutionary terms, more abundant, have a higher probability of being essential, have longer unstructured regions (that are fast evolving), are encoded by genes with more genetic interactions, more protein interactions, and are under tighter post-translational regulation. Shou et al. (Shou, Bhardwaj et al. 2011) recently demonstrated that different types of molecular networks rewire at different rates with their order being (from fast to slow): transcriptional, phosphorylation, genetic

interaction, miRNA, protein interaction and metabolic pathway networks. Therefore, it is conceivable that phosphoproteins act as the raw material for adaption at various evolutionary speeds.

Several of the properties that we observed in the current phosphoproteome were also observed correctly in previous and much smaller datasets, with less stringent filtering criteria. Therefore, despite the incompleteness of the current compendium, we suggest that this high-quality sample is sufficient to accurately reveal the major properties of the entire yeast phosphoproteome.

2.5 Supporting Information

2.5.1 Overlap between any two phosphoproteomic experiments

Distance between any two experiments is calculated with the binary distance function in R. Overlap between any two experiments is calculated as: 1 – binary distance.

	GRU	CHI	LI	ALB	BOD	BEL	HUB	HOLT1	HOLT2	HOLT3	GNAD
CHI	0.96										
LI	0.89	0.93									
ALB	0.94	0.95	0.81								
BOD	0.90	0.93	0.75	0.76							
BEL	0.96	0.98	0.97	0.97	0.97						
HUB	0.95	0.96	0.90	0.94	0.88	0.99					
HOLT1	0.89	0.94	0.72	0.79	0.73	0.97	0.92				
HOLT2	0.93	0.96	0.79	0.79	0.76	0.98	0.94	0.70			
HOLT3	0.94	0.95	0.79	0.79	0.76	0.98	0.95	0.73	0.72		
GNAD	0.93	0.94	0.79	0.84	0.79	0.98	0.91	0.78	0.82	0.83	
SOUFI	0.91	0.92	0.83	0.87	0.83	0.96	0.93	0.84	0.87	0.87	0.79

SI.Table 2.1: Distance matrix for 12HQ p-sites between any two experiments.

	GRU	CHI	LI	ALB	BOD	BEL	HUB	HOLT1	HOLT2	HOLT3	GNAD
CHI	0.80										
LI	0.69	0.77									
ALB	0.78	0.80	0.63								
BOD	0.71	0.75	0.55	0.52							
BEL	0.88	0.90	0.88	0.90	0.89						
HUB	0.87	0.89	0.85	0.90	0.85	0.92					
HOLT1	0.68	0.76	0.52	0.59	0.50	0.89	0.86				
HOLT2	0.73	0.78	0.57	0.51	0.49	0.90	0.89	0.49			
HOLT3	0.74	0.78	0.57	0.50	0.48	0.91	0.89	0.49	0.46		
GNAD	0.75	0.79	0.61	0.67	0.58	0.89	0.86	0.57	0.61	0.62	
SOUFI	0.73	0.75	0.63	0.68	0.63	0.88	0.87	0.62	0.64	0.64	0.59

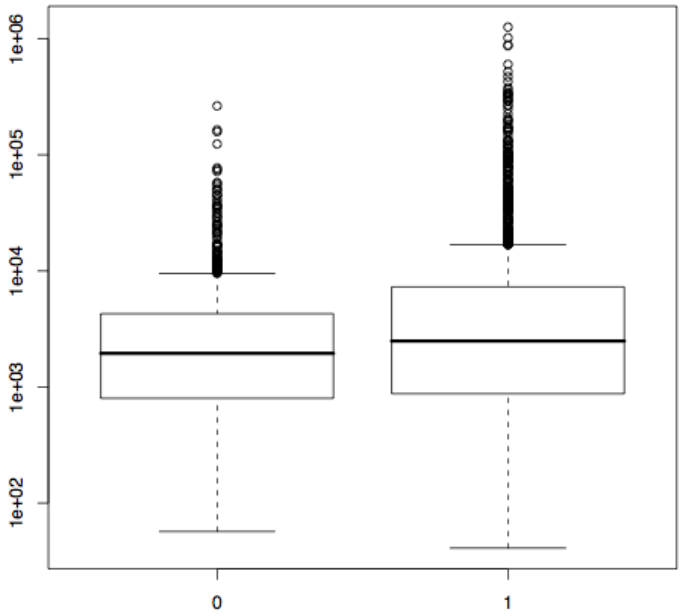
SI.Table 2.2: Distance matrix for 12HQ phosphoproteins between any two experiments.

Table summary: **GRU**: Gruhler et al., 2005; **CHI**: Chi et., 2007; **LI**: Li et al., 2007; **ALB**: Albuquerque et al., 2008; **BOD**: Bodenmiller et al., 2008; **BEL**: Beltrao et al., 2009; **HUB**: Huber et al., 2009; **HOLT1**: Holt et al., 2009 - Asynchronous culture; **HOLT2**: Holt et al., 2009 - Arrested in mitosis with the spindle poison nocodazole; **HOLT3**: Holt et al., 2009 - Arrested in late mitosis by overexpression of a nondegradable cyclin, Clb2-delta; **GNAD**: Gnad et al., 2009; **SOUFI**: Soufi et al., 2009.

2.5.2 Protein abundance

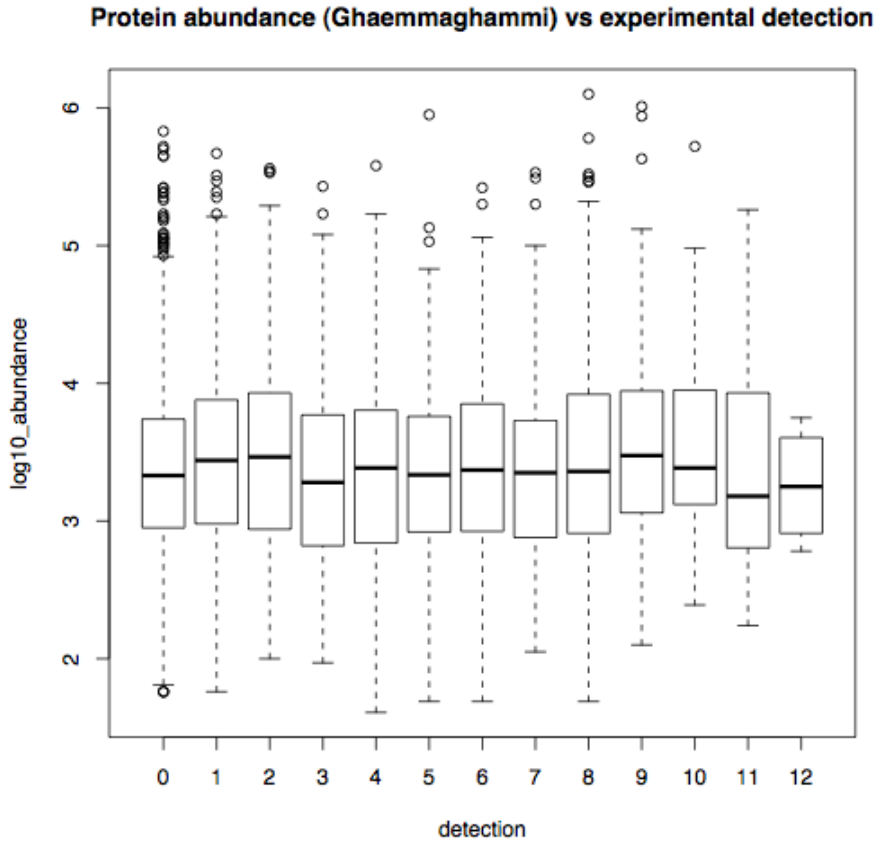
We wanted to investigate if protein abundance could affect the detection of a phosphoprotein in the MS experiments. The abundance of a protein is not a static measurement, but it may change due to, for instance, external stimuli. Therefore, for each HTP MS/MS experiment there should be specific protein abundance measurements, which is not the case at the moment. Therefore, three protein abundance datasets were used, the Ghaemmaghami et al., 2003 for yeast grown to exponential phase in rich medium and the Newman et al., (2006) for yeast grown in rich (YEPD) and synthetic complete (SD) media (Ghaemmaghami, Huh et al. 2003; Newman, Ghaemmaghami et al. 2006). The Ghaemmaghami dataset is well correlated to the other two datasets (Spearman 0.63 and 0.61 for the YEPD and SD respectively), whereas the two Newman datasets are strongly correlated (Spearman 0.91). Given the fact that the difference in protein abundance among two proteins may span 3-5 orders of magnitude (Ghaemmaghami: 41-1,260,000; NewmanYEPD: 45-86,153; Newman SD: 48-57,133), the observed correlations are very good. Therefore, we consider it reasonable (for our purposes) to use these experiments as proxies to extrapolate to the 12 phosphoproteomic datasets that were performed in different conditions. We thus compared protein abundances of the phosphoproteome (2781 ORFs) against the negative phosphoproteome (2219 ORFs) and observed that

the phosphoproteome had, on average, 2-3 times higher abundance than the negative dataset, consistently for each of the three abundance datasets (Wilcoxon p<3e-7). We also looked for each of the 12 experiments, at whether the detected phosphoproteins had a higher abundance (in each of the three datasets) than the rest of the proteome that was not detected. We used all phosphoproteins for each experiment, without applying any filters at all. In 78% of the cases (3 abundance datasets x 12 MS experiments), we found the phosphoproteins to have significantly higher abundance (Wilcoxon p<0.05). Beltrao et al., (2009) also report that phosphopeptides are 3 times more abundant than nonphosphorylated proteins, but given the 8 orders of magnitude span, consider it a small difference (Beltrao, Trinidad et al. 2009). Although a bias clearly exists, we also believe that it is not the determining factor for identifying a phosphoprotein, because there is no clear distinction in the level of abundance of the phosphoproteome vs the non-phosphoproteome. The abundances of both groups of proteins span a similar range of orders of magnitude (see SI.Figure 2.1).



SI.Figure 2.1: Boxplot of protein abundances from the Ghaemmaghami *et al*. Dataset (y axis on logarithmic scale). With 0 we denote the non-phosphoproteome and with 1 the 12HQ phosphoproteome.

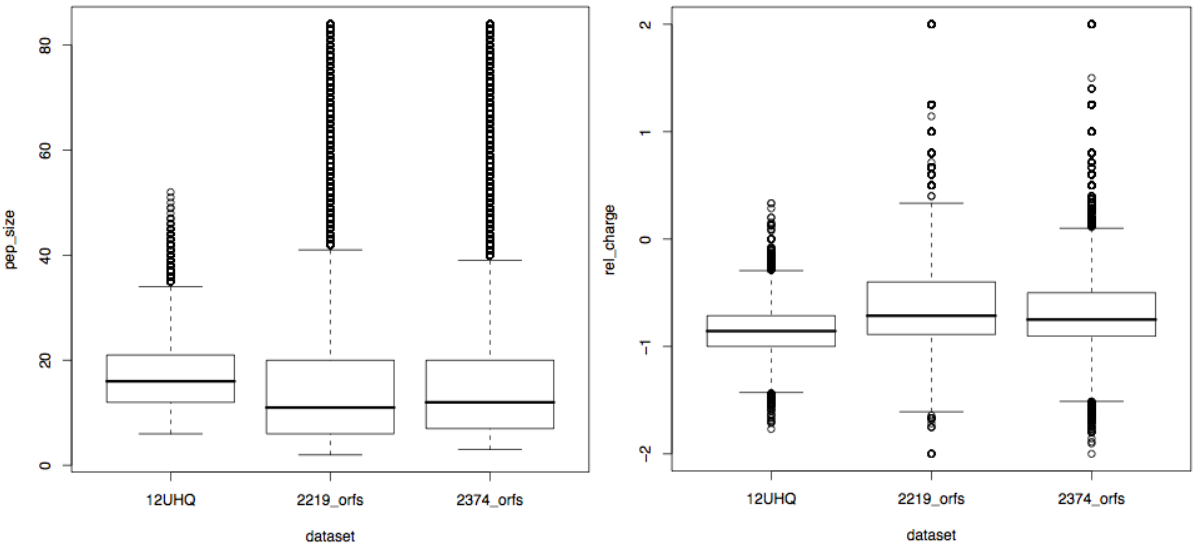
A major concern in our analyses is that a protein is identified as phosphorylated in many experiments, simply because it has a high protein abundance. Therefore, if such an inherent detection bias strongly affects our compendium, one would expect the protein abundance level to increase as the detection level of a protein (number of experiments that was detected as phosphorylated) increases. Nevertheless, as SI.Figure 2.2 shows, our compendium does not suffer from such inherent detectability issues. The same conclusions are reached for the other two protein abundance datasets of Newman et al. (results not shown).



SI.Figure 2.2. Boxplot of protein abundances from the Ghaemmaghami *et al*. Dataset (y axis on logarithmic scale) for the different bins of proteins with certain detection level (number of experiments that was detected as phosphorylated).

2.5.3 Peptide analysis

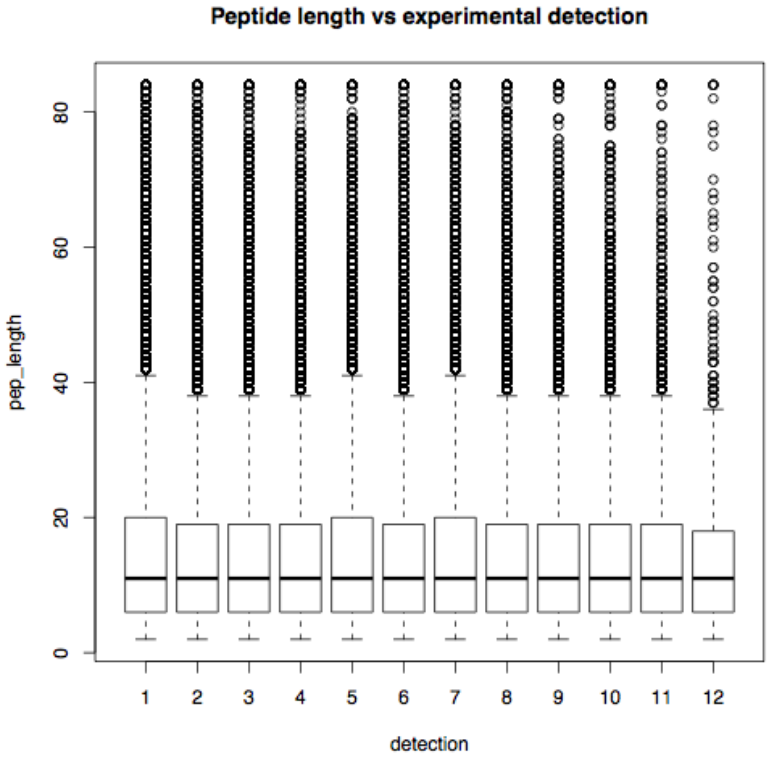
We investigated whether there were likely to be differences in the length or relative charge of the digested peptides generated for MS analysis between phosphorylated and non-phosphorylated proteins. It may be that current protocols can detect a narrow spectrum of the phosphopeptides that are not present in the negative dataset. We performed a theoretical trypsin digestion of the proteins in the two datasets with the proteogest tool (Cagney, Amiri et al. 2003) and calculated the length and relative charge of the theoretical peptides. Although the two datasets were marginally different, we did not observe any significant difference between them (see SI.Figure 2.3).



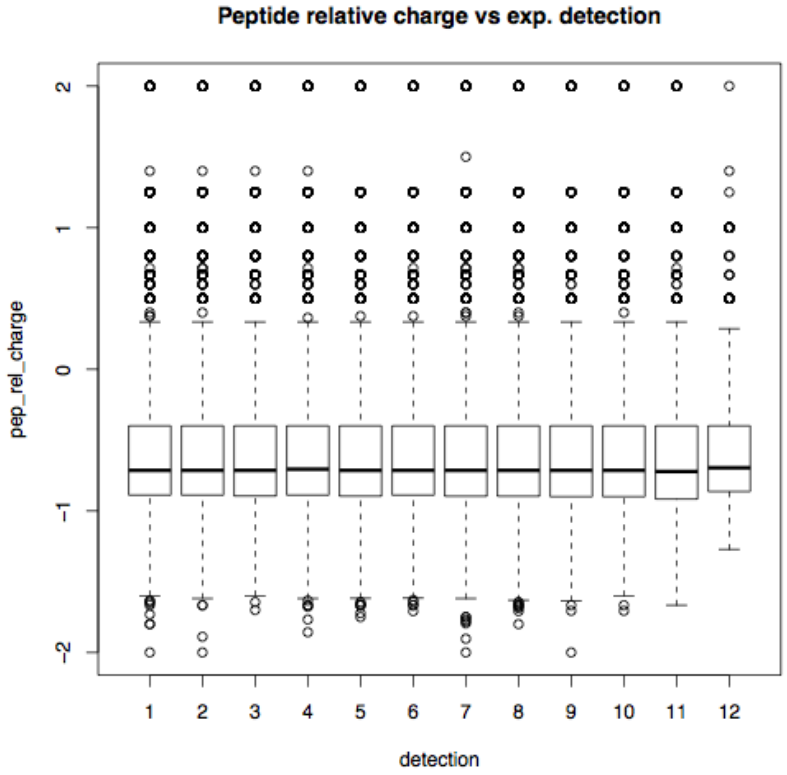
SI.Figure 2.3: Boxplots of the peptide length (A: left) and relative charge (B: right) for the phosphopeptides, for the predicted peptide products of the 2374 12HQ phosphoproteins and the 2219 proteins of the non-phosphorylated set.

Again, a major concern in our analyses is that a protein is identified as phosphorylated in many experiments, simply because its digested peptides have some specific properties, related to peptide length or relative charge. Therefore, if such an inherent detection bias strongly affects our compendium, one would expect the peptide length or relative charge of theoretically digested peptides to be significantly different among the groups of proteins with different detection levels (number of experiments that was detected as phosphorylated). Nevertheless, as

SI.Figure 2.4 & 2.5 show, our compendium does not suffer from such inherent detectability issues.



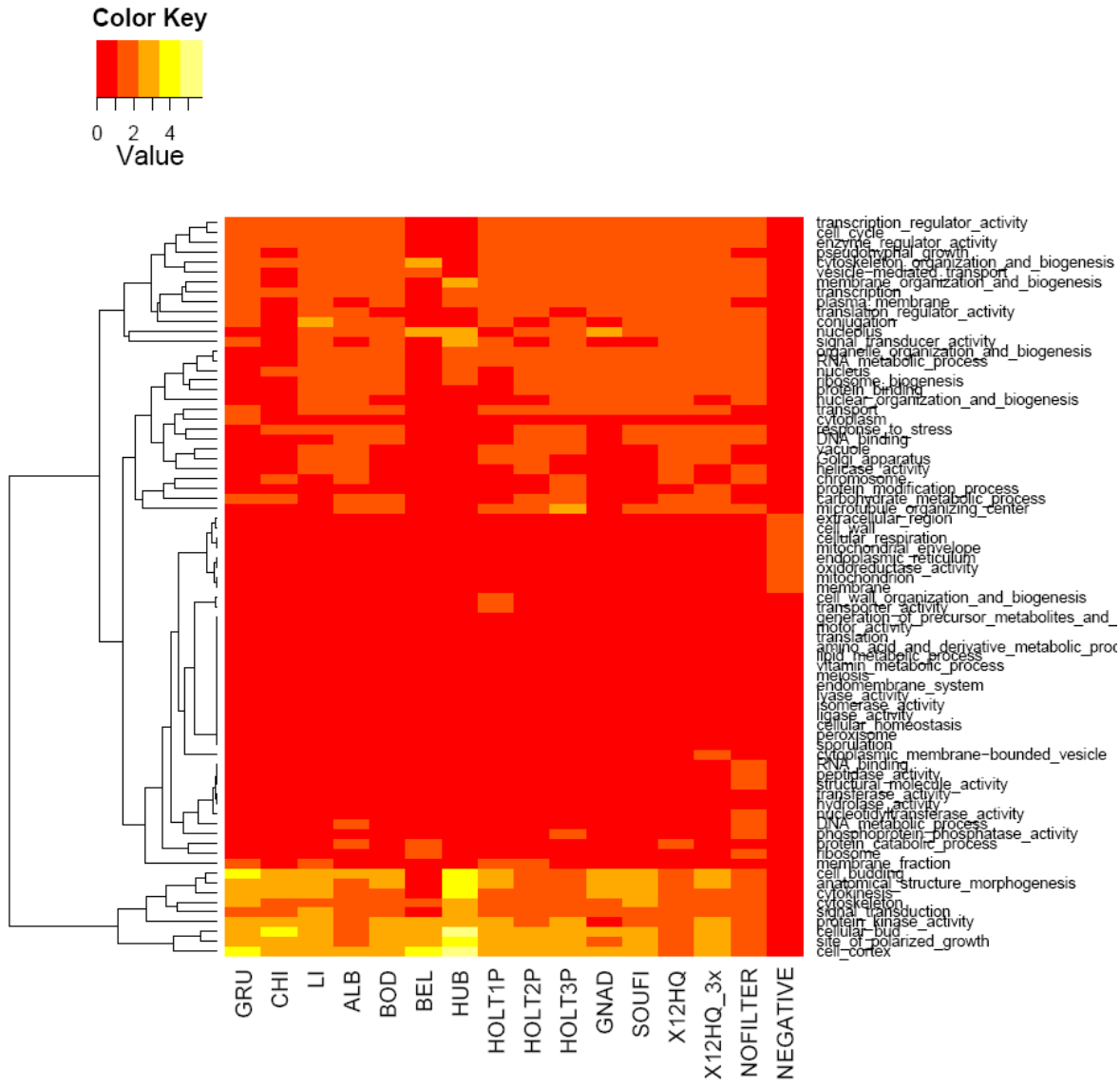
SI.Figure 2.4. Boxplots of the theoretically digested peptide length for the bins of proteins with certain detection level (number of experiments that was detected as phosphorylated).



SI.Figure 2.5. Boxplots of the relative charge of the theoretically digested peptide, for the bins of proteins with certain detection level (number of experiments that was detected as phosphorylated).

2.5.4 GO_Slim heatmap

SI.Figure 2.6 contains a heatmap for the phosphorylated proteins identified in the 12 HTP experiments, constructed using the heatmap.2 function in R. The color denotes the fold-enrichment of the specified functional category.



SI.Figure 2.6: Heatmap for the phosphorylated proteins identified in the 12 HTP experiments. The color denotes the fold-enrichment of the specified functional category.

2.5.5 Controlling for MS-detectability

It is conceivable that the dataset of negative phosphoproteome is an artefact due to its potentially inherent un-detectability by high-throughput (HTP) proteomic Mass Spectrometry (MS) technologies. In order to account for this potentially inherent undetectability, we used two HTP yeast proteomic datasets that were detectable by

MS technology, generated by De Godoy et al., 2008 and by Wu et al., 2011 (de Godoy, Olsen et al. 2008; Wu, Dephoure et al. 2011). These two experiments, combined together identified 4656 proteins. We repeated all the analyses that compare the properties of the 12HQ phosphoproteome against the negative phosphoproteome, but only for those MS-detectable (by De Godoy et al., 2008 or Wu et al., 2011) proteins. By the term “MS-detectable” we refer to those proteins whose peptides may be detected in a high-throughput MS proteomics experiment, without the phosphopeptide enrichment step that is done in phosphoproteomic experiments.

In the De Godoy et al., (2008) experiment, 4399 proteins of haploid and diploid yeast cells, grown in log-phase, were detected with very high accuracy by high-throughput Mass spectrometry (99% certainty about protein identification; 32% average protein coverage by identified peptides). In addition, the data of this experiment overlapped 89% with those of two tagging approaches. The authors also reported no bias against low-abundance or membrane proteins in their dataset. Therefore, this experiment provides a good representation of the yeast proteome.

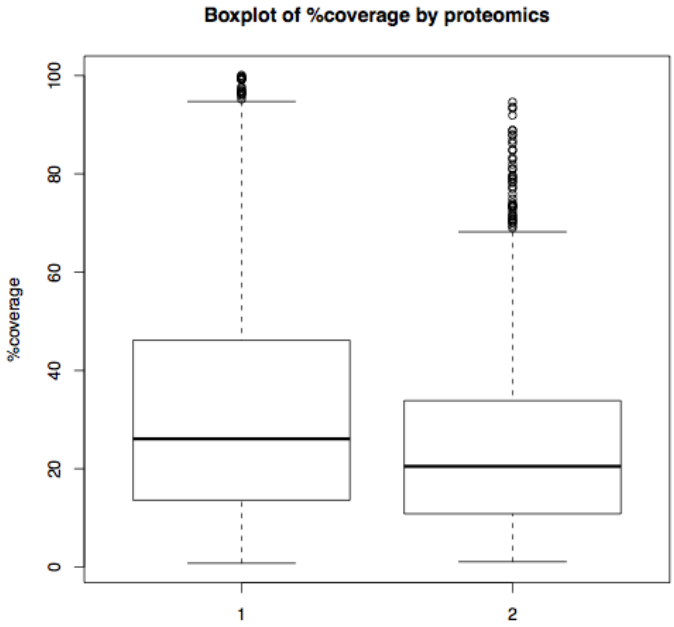
The second experiment of Wu et al., 2011 detected 4281 proteins from three types of yeast cells (wild type, ste7Δ & fus3Δ) grown until they reached log-phase. This second dataset of Wu et al., 2011 is very similar in detected protein content with the De Godoy et al., 2008 experiment. The combination of the two datasets (De Godoy et al., 2008 and Wu et al., 2011) yields 4656 yeast proteins, where 86% of them are found in both datasets, thus confirming the reproducibility of the MS technology for protein detection, even by different laboratories.

The first observation was that 94% of the 12HQ phosphoproteome (2239/2374 proteins) compared to 64% of the negative-phosphoproteome (1418/2219 proteins) are MS-detectable in that particular condition (log-phase). We designate those 2239 MS-detectable phosphoproteins as MS_D_Ph and the 1418 MS-detectable non-phosphoproteins as MS_D_nonPh. Although at first sight this difference in

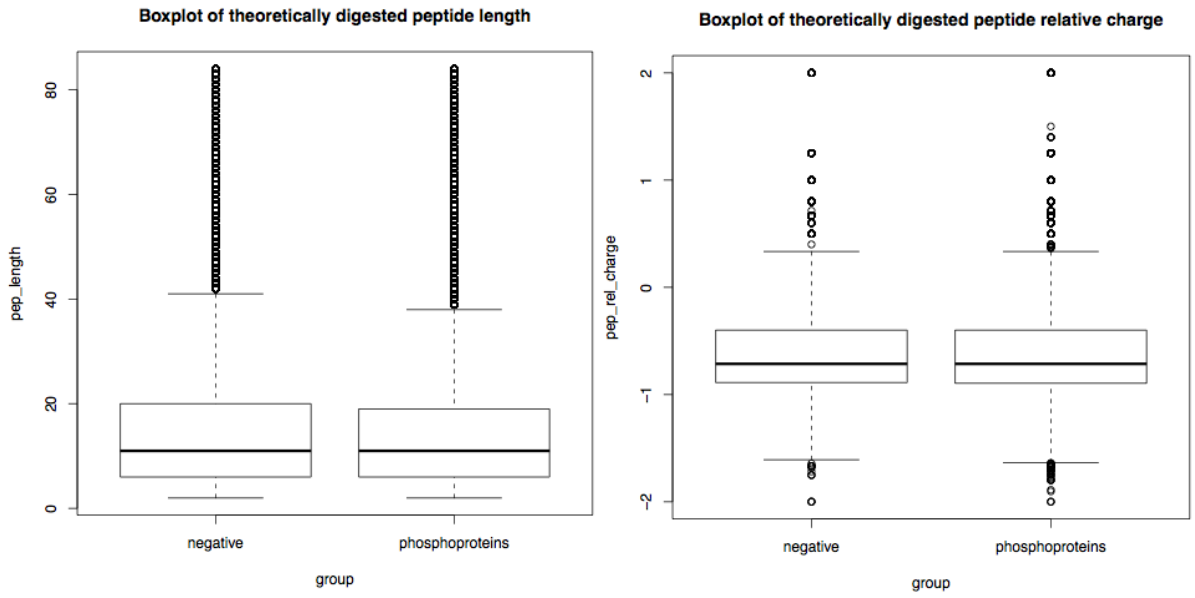
detectability observed above (94% vs 64%) could be a reason for concern, our analyses further on clearly demonstrate that the conclusions drawn (regarding the differences between phosphoproteins and non-phosphoproteins) are robust, even when accounting for MS-detectability. In addition, we cannot exclude the possibility that the rest 36% (801/2219 proteins) of the non-phosphoproteome that was not detected in any of these two experiments is actually MS-detectable, but under some other biological conditions. Therefore, this type of MS-detectability criterion that we applied constitutes a very stringent filter.

The second observation was that the MS-detectable phosphoproteins had an average peptide coverage of 32%, compared to 25% for the MS-detectable non-phosphoproteome (see SI.Figure 2.7). For this analysis, the estimated peptide coverage, provided from the de Godoy et al., 2008 dataset was used. Boxplot analysis further demonstrates that although MS-detectable phosphoproteins are more covered by MS peptides than the MS-detectable non-phosphoproteome, the observed difference is not so great so as to justify the hypothesis that these MS-detectable non-phosphorylated proteins cannot be detected as phosphorylated in any phosphoproteomic experiment, due to very low peptide coverage.

The theoretically digested (with trypsin) peptide length and peptide relative charge of the phosphoproteins and the non-phosphoproteins is very similar, as observed in the boxplots in SI.Figure 2.8. For this analysis, the MS-detected proteins from the de Godoy et al., 2008 and Wu et al., 2011 datasets were used.



SI.Figure 2.7: Boxplot of % coverage of proteins by identified peptides in de Godoy et al., 2008. The group of MS-detectable phosphoproteins is one the left (designated with 1), whereas the group of MS-detectable non-phosphoproteins is on the right (designated with 2).



SI.Figure 2.8: Boxplots of the peptide length (A: left side) and relative charge (B: right side) for the phosphopeptides, for the predicted peptide products of the MS-detectable 12HQ phosphoproteins and the MS-detectable proteins of the non-phosphorylated set.

The enrichment of GO-slim categories is similar for the original negative phosphoproteome (2219 proteins; designated as “Negative” in the GO-Slim image)

and the MS-detectable negative phosphoproteome (1418 proteins; designated as “Negative2” in the GO-Slim image).

Regarding the differences observed between phosphorylated and non-phosphorylated proteins, we reach the same conclusions even when we control for MS-detectability. That is:

- Phosphoproteins have (on average), 53% shorter protein half-lives (Wilcoxon p<0.0004) than non-phosphorylated proteins (for this analysis we used only genes with protein half-life measurements).
- A higher fraction of phosphoproteins are ubiquitinated, compared to the non-phosphoproteome (28% vs 14% respectively; Chi-squared p<3e-16).
- A higher fraction of phosphoprotein genes are essential, compared to the genes of the non-phosphoproteome (24% vs 15% respectively; Chi-squared p<6e-10).
- Phosphoprotein genes have, on average, 18-22% more genetic interactions than the non-phosphorylated dataset (Wilcoxon p<0.0005),
- phosphoproteins have on average 33-110% more protein-protein interactions (Wilcoxon p<2e-9) than non-phosphoproteins (the fluctuation of this percentage depends on whether we use only proteins for which there is at least one protein-interaction, or not).
- Phosphoproteins have on average 83-204% more interactions with kinases (Wilcoxon p<2e-12) than non-phosphoproteins (the fluctuation of this percentage depends on whether we use only proteins for which there is at least one kinase-interaction or not).
- Phosphoproteins have on average 184% longer intrinsically disordered (ID) regions (Wilcoxon p<8e-199) than non-phosphoproteins. Furthermore, phosphoproteins have on average 25% longer non-ID regions (Wilcoxon p<3e-7) than non-phosphoproteins.
- Phosphoproteins have on average 153-224% higher protein abundance (Wilcoxon p<8e-5) than non-phosphoproteins (the fluctuation of this percentage depends on which of the three protein abundance datasets we

use), in accordance with the original observation where no MS-detectability filtering was applied. Furthermore, boxplot analyses clearly demonstrate that the two groups span similar orders of magnitude.

- There is no significant difference in the number of TFs that bind at the promoters of phosphoprotein or non-phosphoprotein genes (for this analysis we used only genes with at least one TF binding at their promoter).

A dataset of fungal orthologous groups by Wapinski et al., 2007 (Wapinski, Pfeffer et al. 2007) was used in order to identify yeast orthologs in other fungal genomes (this dataset incorporates the pillars from Yeast Gene order browser (Byrne and Wolfe 2005). Next, the percentage of yeast genes with orthologs in another species was calculated for both the phosphoproteins and the negative phosphoproteome group and their difference as a ratio was estimated, together with the statistical significance. This comparison was performed for all proteins (blue bars in Figure 2.6), for MS-detectable only proteins (red bars in figure 6 of main text) and for proteins that were not detected by HTP-MS (yellow bars in figure 6 of main text). A ratio above 1 (that was observed in all comparisons) indicates that phosphoproteins have more orthologs than the negative phosphoproteins, in any given fungal species. For all comparisons but two (MS-detectable proteins in *D. hansenii* and *Y. lipolytica*), the difference is statistically significant (chi-squared $p<0.05$).

2.6 Author contributions

Y.H. helped prepare the phosphorylation dataset and conducted part of bioinformatics analysis under the supervision of G.A. and Y.V.D.P..

3 Post-translational regulation impacts the fate of duplicated genes

Amoutzias, G*., He, Y*., Gordon, J., Mossialos, D., Oliver, S., Van de Peer, Y.

Post-translational regulation impacts the fate of duplicated genes.

Redrafted from *Proc Natl Acad Sci* 107, 2967-2971. (*equal contribution)

Abstract

Gene and genome duplications create novel genetic material on which evolution can work and have therefore been recognized as a major source of innovation for many eukaryotic lineages. Following duplication, the most likely fate is gene loss; however, a considerable fraction of duplicated genes survive. Not all genes have the same probability of survival, but it is not fully understood what evolutionary forces determine the pattern of gene retention. Here, we use genome sequence data as well as large-scale phosphoproteomics data from the baker's yeast *Saccharomyces cerevisiae*, which underwent a whole-genome duplication ~100 mya, and show that the number of phosphorylation sites on the proteins they encode is a major determinant of gene retention. Protein phosphorylation motifs are short amino acid sequences that are usually embedded within unstructured and rapidly evolving protein regions. Reciprocal loss of those ancestral sites and the gain of new ones are major drivers in the retention of the two surviving duplicates and in their acquisition of distinct functions. This way, small changes in the sequences of unstructured regions in proteins can contribute to the rapid rewiring and adaptation of regulatory networks.

3.1 Introduction

Whole-genome duplications (WGDs) have occurred repeatedly in eukaryotic evolution and have been linked to genetic innovation, adaptation, speciation and survival (Ohno 1970; Freeling and Thomas 2006; Scannell, Byrne et al. 2006; Fawcett, Maere et al. 2009). Following a WGD, most duplicate copies are lost (Lynch and Conery 2000; Maere, De Bodt et al. 2005), but a considerable fraction survive, with either selection or genetic drift accounting for the pattern of duplicate gene retention. Interestingly, this pattern of retention is not random, but, rather, biased to certain functional categories. In particular, genes involved in regulation are preferentially retained (Davis and Petrov 2005; Maere, De Bodt et al. 2005; Freeling and Thomas 2006) and it is this preferential retention that likely predetermines the future of a lineage. However, the mechanisms that determine which genes are maintained in duplicate and which return to a single-copy state, are largely unknown.

After a WGD, there is a relatively short period of genome instability, extensive gene loss and elevated levels of nucleotide substitution (Otto 2007). During that period, regulatory networks must be rapidly re-wired to integrate the newly duplicated (and, at the same time, diverging) genes and thus prevent chaos in the control of cellular processes. Rapid evolution and functional divergence has indeed been observed at the level of the transcription of duplicated genes (Li, Yang et al. 2005; Casneuf, De Bodt et al. 2006), which is usually explained by point mutations in short transcription factor binding motifs. However, since the effectors of gene action are proteins, adaptation might also occur at the post-translational level of regulation. Since the amino-acid sequence motifs for post-translational modification (PTM), and especially phosphorylation, are short (Gnad, Ren et al. 2007) and occur within rapidly evolving unstructured regions (Iakoucheva, Radivojac et al. 2004), we reasoned that changes in PTM sites might present a ready means of rapidly effecting the necessary re-wiring. Furthermore, we wanted to explore whether the rapid evolution of these sites might

be linked to gene retention. Therefore, we exploited the wealth of proteomic and genomic data available for the baker's yeast *S. cerevisiae* and examined the relationship between protein phosphorylation, gene retention, and functional divergence following the WGD that occurred in the hemiascomycete yeasts, about 100 mya (Wolfe and Shields 1997; Kellis, Birren et al. 2004).

3.2 Results and Discussion

3.2.1 Retained duplicates are highly phosphorylated

There have been several large-scale *in vivo* studies (Gruhler, Olsen et al. 2005; Chi, Huttenhower et al. 2007; Li, Gerber et al. 2007; Reinders, Wagner et al. 2007; Albuquerque, Smolka et al. 2008; Bodenmiller, Campbell et al. 2008) of the phosphoproteome of *S. cerevisiae,* using highly reproducible techniques of mass spectrometric analysis. Since the identified phosphopeptides could match one or more open reading-frames (ORFs), we generated two phosphorylation datasets. One contains phosphopeptides that have a unique and exact match (designated 6eU; 6 refers to the six datasets used), the other contains phosphopeptides that exactly match more than one ORF (designated 6eNU) (section 3.5.1). All subsequent analyses were performed on both these datasets (whenever applicable), and the same conclusions were obtained from both (statistical significance through Wilcoxon tests, data not shown). The datasets compiled from these six studies indicate that 8,500-11,300 phosphorylation sites (p-sites) are distributed over 2,200-2,400 proteins in *S.cerevisiae*. GO_slim analysis confirmed that these datasets were enriched for proteins localized in the nucleus, and involved in signal transduction and transcription regulation (section 3.5.3).

Previous analysis of the *S.cerevisiae* genome (Byrne and Wolfe 2005) identified over 500 gene pairs that were produced by the WGD (which we designate WGD-genes or ohnologs, and their products WGD-proteins) and ~4000 genes that were duplicated in the WGD but later returned to single-copy status (RSS-genes, RSS-proteins) (section 3.5.4). Of note, these RSS genes constitute most of the remainder of the *S. cerevisiae* genome after the WGD-produced duplicate pairs are excluded. We found that a higher fraction of WGD-genes encoded phosphoproteins compared to the RSS-genes (48-58% *vs.* 42-43%), a statistically significant difference ($p<1e^{-3}$, χ^2 test). Furthermore, phosphoproteins of the WGD-group have, on average, significantly more p-sites than those of the RSS-group (4.6 and 3.5 sites per protein, on average, for 6eU; $p<1.64e^{-7}$, Wilcoxon). This observation is supported not only from experimental data, but also from *in silico*-predicted data (see section 3.5.5; $p<3.4e^{-14}$, Wilcoxon). Furthermore, this observation is robust with respect to the selection, quality, and evolution of the various experimental datasets (see section 3.5.6). A jackknife analysis considering all six experimental datasets showed that, no matter which dataset is excluded from the analysis, WGD-phosphoproteins have on average more p-sites (see section 3.5.6; $p<0.0013$, Wilcoxon).

It is possible that this observation is very significant, but is not general, being confined to only a few gene categories. In fact, GO_slim analysis showed that the enrichment of p-sites in WGD-phosphoproteins *vs.* RSS-phosphoproteins is statistically significant (see section 3.5.7) for one-third (6eU) to two-thirds (6eNU) of the GO_slim categories. Only for one category, namely structural molecule activity, in one dataset, did we observe the inverse trend (see Table 3.9). A jackknife analysis, considering all GO_slim categories, showed that, no matter which category was removed, the WGD group still contained more p-sites than the RSS group (see section 3.5.7; $p <1e^{-4}$, Wilcoxon). Signaling and transcription factor (TF) molecules show higher retention than average after a WGD event and also show higher levels of phosphorylation than average. In order to ensure that the latter observation is not a trivial consequence of the large number of TFs and signaling molecules in our dataset, we removed all TFs, kinases, phosphatases and cyclins and still found the

WGD phosphoproteins to contain more p-sites than the RSS phosphoproteins (see section 3.5.7; $p < 1.5e^{-6}$, Wilcoxon). Furthermore, in order to account for gene dosage imbalances (Papp, Pal et al. 2003; Mintseris and Weng 2005), we also removed i) all the ribosomal proteins or ii) all known protein complexes and found that our conclusions are still robust (see section 3.5.7; $p<2.2e^{-9}$, Wilcoxon and section 3.5.7; $p<9.8\ e^{-7}$, Wilcoxon, respectively). We also controlled for potential biases arising from i) differences in protein abundance (Ghaemmaghami, Huh et al. 2003) (see section 3.5.8; $p < 3.3\ e^{-5}$, Wilcoxon) or ii) coverage (see section 3.5.8; p=0.009, Wilcoxon) in the various experiments, iii) essentiality of genes (Giaever, Chu et al. 2002; Steinmetz, Scharfe et al. 2002; Pereira-Leal, Audit et al. 2005; Zotenko, Mestre et al. 2008; Pache, Babu et al. 2009) (see section 3.5.8; p<2.3e-6, Wilcoxon), or iv) protein interaction network centrality (Pereira-Leal, Audit et al. 2005; Batada, Reguly et al. 2006; Zotenko, Mestre et al. 2008) (see section 3.5.8; $p<3e^{-7}$, Wilcoxon) and still found that WGD-proteins contained more p-sites. The importance of taking into consideration protein function in evolutionary analyses as performed here has been highlighted previously (Kunin, Pereira-Leal et al. 2004).

3.2.2 Inference of ancestral phosphorylation sites in the pre-WGD ancestor

Based on known *S. cerevisiae* p-sites and multiple sequence alignments of orthologs in three species that diverged from *S. cerevisiae* just prior to the WGD event (Figure 3.1A), namely *Ashbya* (*Eremothecium*) g*ossypii* (Dietrich, Voegeli et al. 2004)*, Kluyveromyces lactis* (Dujon, Sherman et al. 2004) and *Kluyveromyces waltii* (Kellis, Birren et al. 2004), we inferred the presence of ancestral p-sites in the proteins of the pre-WGD ancestor (see section 3.5.9). To achieve this, we used both the 6eU and 6eNU datasets and, by applying various levels of stringency based on the variation of amino acids surrounding the p-site (see section 3.5.9), we generated eight different sets of ancestral p-sites (see Materials and Methods). In all eight sets, we again observed that, on average, the ancestral proteins of the WGD-group contained

significantly more p-sites than ancestral RSS-proteins (see section 3.5.9; $p<3.4\ e^{-9}$, Wilcoxon). Since we used the ORFs of both retained duplicates to infer the ancestral p-sites of WGD-proteins and only one RSS-ORF to infer the ancestral p-sites of RSS-proteins, it is possible that the observed difference reflects this bias. Therefore, we repeated the analysis using only one of the two ohnologs (the one with most p-sites) to infer the ancestral p-sites. Although we consider this already stringent, since we underestimate the p-sites of the ancestral molecule that later gave rise to sub-functionalized copies, see further, we still observed that ancestral proteins of the WGD-group contained significantly more p-sites (see section 3.5.9; $p<8.4\ e^{-5}$, Wilcoxon). Recently, concerns have been raised about the possibility that many p-sites are not functional (Lienhard 2008; Landry, Levy et al. 2009). However, since our evolutionary analysis is based on p-sites that have been conserved for more than 100 million years, we believe that there is little chance that our conclusions are a trivial consequence of an accumulation of non-functional p-sites in WGD duplicates.

To see whether the link between the number of ancestral p-sites and gene duplicate retention is a general phenomenon, and not just confined to *S.cerevisiae*, we examined the gene retention patterns in three more post-WGD species (*Candida glabrata* (Dujon, Sherman et al. 2004)*, S.castellii* (Dujon, Sherman et al. 2004)*,* and *Kluyveromyces polysporus* (Scannell, Byrne et al. 2006)). We are aware of the fact that the evolutionary process is not entirely independent in all these species, since they share a common ancestor. However, they diverged from each other shortly after the WGD event (Scannell, Byrne et al. 2006). Products of genes that survived as duplicates in the genomes of at least three of the four post-WGD yeast species are designated ‘retained-in-majority’, whereas products of genes that have returned to single-copy status in the genomes of at least three of those four post-WGD species are designated ‘lost-in-majority’. For all eight ancestral p-site datasets (see Materials & Methods), the pre-WGD ancestors of the ‘retained-in-majority’ category had, on average, more p-sites than the pre-WGD ancestors of the ‘lost-in-majority’ category, meaning that proteins with more p-sites have repeatedly been retained in duplicate,

compared to proteins with fewer p-sites (see section 3.5.9; $p<2.2e^{-8}$, Wilcoxon) (Figure 3.1B).

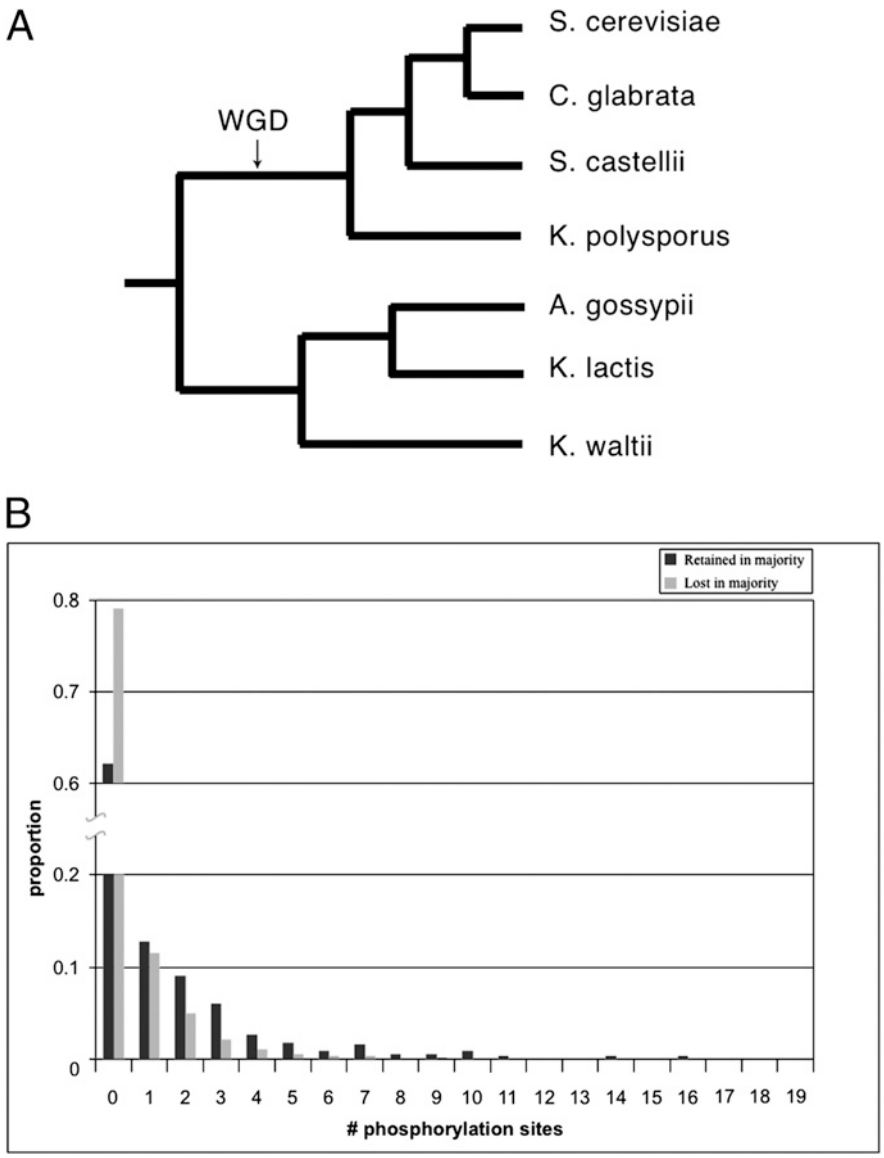


Figure 3.1: The number of protein phosphorylation sites in the pre-WGD ancestor affects the probability of gene retention in the four post-WGD yeast lineages. (*A*) Species tree showing the evolutionary relationships of the yeast species discussed. (*B*) The number of ancestral p-sites affects the fate of duplicate retention in four post-WGD lineages (*S. cerevisiae*, *C. glabrata*, *S. castellii*, and *K. polysporus*). The graph shows that the ancestors of the "retained-in-majority" bins had more p-sites than the ancestors of the "lost-in-majority" bins (see Supporting Information for details).

3.2.3 Sub- and neo-functionalization of phosphorylation sites

How might protein phosphorylation affect the retention of duplicates? Previous analyses have suggested that the partitioning of functions (sub-functionalization) between the two copies or the emergence of new functions (neo-functionalization) for one or both copies of a duplicated gene favors their retention (see Figure 3.2). In order to measure the effect of phosphorylation-related sub-functionalization, we assumed that such a partition had occurred if duplicates lost a complementary set of p-sites. We used all the pre-WGD genes (ancestors of both WGD- and RSS-genes) whose proteins had at least two ancestral p-sites and measured how many of these ancestral genes duplicated to give pairs with signs of sub-functionalization (see section 3.5.9). We found that between 2.5% and 7% (depending on dataset) of those ancestral genes gave rise to sub-functionalized copies in *S. cerevisiae*. For example, 5-12 (depending on stringency of criteria to infer ancestry) ancestral p-sites underwent sub-functionalization in the *BOI1*/YBL085W-*BOI2*/YER114C WGD pair; these proteins are involved in bud emergence and polar growth. In addition, we observed that ancestral proteins whose ohnologs sub-functionalized had, on average, more p-sites than all other ancestral proteins (see section3.5.9; $p<1.4\ e^{-4}$, Wilcoxon), and also had more p-sites than those retained in duplicate without sub-functionalization (see section 3.5.9; $p<0.011$, Wilcoxon). This finding supports the idea of a stochastic process of reciprocal loss of functional p-sites. The more p-sites in the ancestral protein, the greater the chance of reciprocal loss in future duplicates, thus leading to sub-functionalization and retention; a pattern in accordance with a model proposed for regulatory sequences (Lynch and Conery 2000). It should be noted that the current phosphorylation dataset is incomplete and, as more phosphorylation data are generated, the number of sub-functionalization cases is likely to increase.

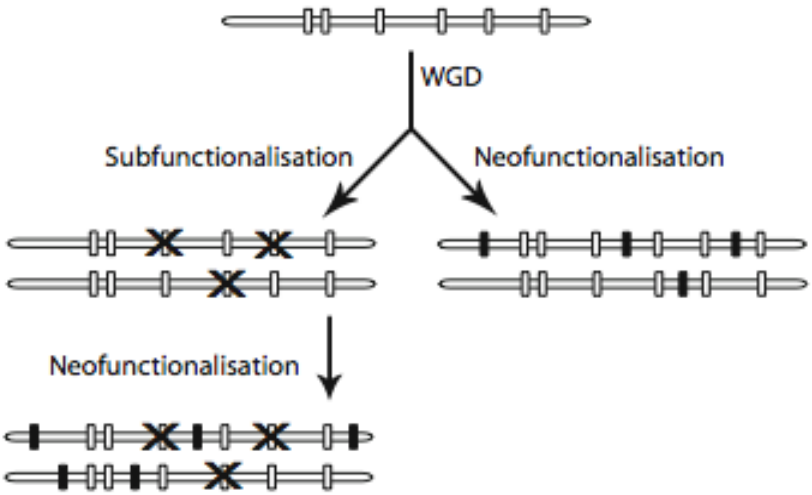


Figure 3.2: Protein phosphorylation and the retention of duplicated genes. Sub-functionalization is the partitioning of p-sites among the duplicates via point mutations and stochastic reciprocal loss. A parallel or (more likely) subsequent event is neo-functionalization, the emergence of new p-sites, again via point mutations. Open boxes refer to ancestral p-sites, and solid boxes refer to recently emerged p-sites.

We also identified potential cases of neo-functionalization by looking for p-sites not present in the pre-WGD ancestor (see section 3.5.9). We refer to such sites as neo-p-sites and assume that some new regulatory interaction may have evolved. Without mutation data, we do not know whether these neo-p-sites are truly functional (Lienhard 2008). Furthermore, without equally extensive phosphorylation data from other yeasts we cannot take into account those p-sites that have undergone evolutionary turnover (Holt, Tuch et al. 2009), where a functional p-site is lost, but a new neighboring p-site emerges and rescues its function. Thus, we might overestimate the significance of neo-functionalization in yeast ohnologs; on the other hand, the current phosphorylation dataset is incomplete. Nevertheless, 29-40% of ohnologs seemed to have acquired one or more novel p-sites; moreover, 73-94% of ohnologs that undergo sub-functionalization simultaneously seem to undergo neo-functionalization, which is in agreement with a complex model of neo-sub-functionalization (He and Zhang 2005). The high incidence of novel p-sites is in accordance with previous reports on the importance of neo-functionalization in TF regulatory motifs (Tirosh and Barkai 2007). Over 80% of p-sites are found within unstructured and fast-evolving loops that comprise ~55% of the protein length (see section 3.5.10) and these regions are linked to tight regulation (Gsponer, Futschik et al. 2008). We observed a significant correlation (Pearson coefficient, 0.44-0.45)

between the absolute length of the unstructured loops and their number of p-sites. The retained duplicates encode phosphoproteins with loops that are, on average, 14% longer than those of RSS-proteins. To see whether the higher incidence of p-sites, and therefore retention, was affected by WGD-protein loops being longer, we normalized our phosphorylation data for loop length and confirmed again that WGD-proteins have more p-sites than RSS-proteins ($p<0.0083$, Wilcoxon). We repeated the same analysis for intrinsic disorder and again noticed that WGD-proteins have more p-sites than RSS-proteins ($p<8.5e^{-5}$, Wilcoxon).

3.2.4 Post-translational modifications in general and not only phosphorylation likely affect the retention of duplicated genes

As we have shown here, WGD-proteins are subject to more phosphorylation than RSS-proteins. While the only extensive *in-vivo* data on PTM concern protein phosphorylation, there are indications that other PTMs are linked to increased levels of retention following duplication (see section 3.5.11). A higher fraction of WGD-proteins than RSS-proteins are ubiquitinated (23.5% *vs.* 19.5%; $p<0.004$, χ^2 test); furthermore WGD-proteins seem to have shorter half-lives than RSS-proteins (see section 3.5.11; $p<7\ e^{-4}$, Wilcoxon). All these results are congruent with our hypothesis that changes in post-translational modification represent rapid and facile routes to the sub- and neo-functionalization of duplicated genes following WGD and thereby promote the retention of duplicate pairs, although they do not explain all cases of duplicate retention (i.e. selection for higher dosage for genes encoding ribosomal proteins).

The impact of phosphorylation on gene retention is probably not confined to WGD, but to small-scale gene duplication (SSD) as well. Several studies have shown that the mode of duplication (WGD vs SSD) has different effects on the evolution of the genome (Davis and Petrov 2005; Guan, Dunham et al. 2007; Hakes, Pinney et al.

2007). Since SSDs occur continuously and at various times, it is not possible to repeat the evolutionary analysis that was possible for the WGD. Nevertheless, when we compared properties of SSD *versus* singleton proteins (see section 3.5.12), we observed that: i) a higher fraction of SSD-proteins are phosphorylated (42-43% *vs*. 33-34%; $p<4\ e^{-9}$, χ^2 test), ii) a higher fraction of SSD-proteins are ubiquitinated (20% *vs*. 14%; $p<2.8\ e^{-9}$, χ^2 test), iii) SSD phosphoproteins have, on average, more p-sites than singleton phosphoproteins ($p<3.2e^{-4}$, Wilcoxon), and iv) SSD-proteins have shorter half-lives ($p<0.0455$, Wilcoxon). Experimental data (Wilson-Grady, Villen et al. 2008) from *Schizosaccharomyces pombe*, a very distant relative of *S.cerevisiae* that did not undergo a WGD, also shows a higher fraction of SSD-proteins being phosphorylated compared to singletons (22.5% *vs*. 14%; $p<2e^{-14}$, chi-squares), although the paucity of functional data for this species limits the analysis (see section 3.5.13).

The higher level of phosphorylation observed, not only for WGDs but also for SSDs, seems to imply that the retention of highly phosphorylated proteins in the yeast lineages cannot directly be attributed to stoichiometric constraints. According to the dosage balance hypothesis (Papp, Pal et al. 2003; Freeling and Thomas 2006; Birchler and Veitia 2007), it is conceivable that proteins involved in phosphorylation might need to maintain relative stoichiometry with their kinases, thus promoting co-retention. Therefore, retention of these highly phosphorylated proteins, if due to stoichiometric balances, should be favored only after a WGD event and not after SSD events. Further research is necessary to see why this does not seem to be the case.

3.3 Conclusions

It is clear from this study that proteins retained in duplicate are subject to more post-translational control and particularly to more phosphorylation, than RSS-proteins. Post-translational regulation repeatedly affected the future of gene duplicates in the

various post-WGD yeast lineages, suggesting that gene retention is, to some extent, pre-determined. The evolutionary analyses performed are congruent with our hypothesis that changes in post-translational modification represent rapid and facile routes to the sub- and neo-functionalization of duplicated genes and thereby promote the retention of duplicate pairs. Our observation is also in accordance with previous observations that 'complex' genes, where complexity is defined by the number of protein domains encoded and *cis*-regulatory elements, tend to be retained more frequently (He and Zhang 2005). An alternative explanation (that does not exclude the previous one) is that tighter regulatory control can buffer the slightly deleterious mutations of duplicated copies that are under relaxed selection and thus provide them with more time to explore the fitness landscape. It may be, after all, that the cell does not favor the survival (for a long time) of a degenerate gene copy that can act like a 'loose cannon'.

3.4 Materials and Methods

3.4.1 Phosphorylation data

For *S. cerevisiae*, we used six publicly available experimental data sets (Gruhler, Olsen et al. 2005; Chi, Huttenhower et al. 2007; Li, Gerber et al. 2007; Reinders, Wagner et al. 2007; Albuquerque, Smolka et al. 2008; Bodenmiller, Campbell et al. 2008), while for every one of them the filter proposed in that specific study for identifying the exact location of phosphorylation sites was applied. All those experiments rely on affinity-based methods (IMAC) for phosphopeptide isolation and their results are estimated to be up to 93% reproducible (Bodenmiller, Mueller et al. 2007). The identified phosphopeptides could match (exactly) either one or more than one ORF. Accordingly, we generated two phosphorylation datasets: one that contains phosphopeptides that have a unique match (designated as 6eU), and one that contains phosphopeptides that may exactly match one or more ORFs

(designated as 6eNU). The reason is that this data treatment may have an effect on the analyses of gene duplication. Transposable elements were removed from the two datasets. For *Sc. pombe*, we used one publicly available phosphoproteomics experiment (Wilson-Grady, Villen et al. 2008).

The NetPhosYeast software (Ingrell, Miller et al. 2007) was used to predict protein phosphorylation sites for the entire *S. cerevisiae* proteome. The NetPhosYeast software is specifically tailored to *S. cerevisiae* and has been shown to outperform all other predictors. We ran the predictions with two different stringent cut-offs, 0.75 and 0.85.

3.4.2 Statistics and Gene Ontology analyses

The R programming language was used for statistical analyses. For Gene Ontology related analyses, we used the GO-slim annotation of yeast (Ashburner, Ball et al. 2000).

3.4.3 Duplication datasets for *S. cerevisiae*

We analyzed genes of *S. cerevisiae*, a species that underwent a WGD ~100 mya (Byrne and Wolfe 2005). Genes were separated into those 1096 that were retained in duplicate following the WGD until today (designated as WGD-genes) and those 4002 that underwent the duplication but later lost the duplicate (designated as Returned to Single Status: RSS-genes). The assignment of orthologs and ohnologs (gene duplicates resulting from a WGD) was based on syntenic information (Byrne and Wolfe 2005), using the *S. cerevisiae* genome (Goffeau, Barrell et al. 1996) as well as

genomes that did not undergo the WGD. For the small-scale gene duplication analysis, we first removed the 1096 WGD genes from the original list of the 5795 protein-coding genes. Next, we defined 2439 *S. cerevisiae* genes as singletons, based on the fact that their proteins had no Blast-p hit against any other *S. cerevisiae* proteins, at a cut-off level of 1e-3. As Small-Scale gene Duplicates (SSDs), we identified 2260 genes that belong neither to the WGD group, nor to the Singletons group.

3.4.4 Inferring the phosphorylation sites of the pre-WGD ancestor

Based on orthology-paralogy relationships identified previously (Byrne and Wolfe 2005), we aligned each *S. cerevisiae* protein (and ohnolog, whenever appropriate) with its orthologs from the three pre-WGD species (Dietrich, Voegeli et al. 2004; Dujon, Sherman et al. 2004; Kellis, Birren et al. 2004) (*Ashbya [Eremothecium] gossypii, Kluyveromyces lactis, Kluyveromyces waltii*), using the t-coffee software (with default parameters) (Notredame, Higgins et al. 2000).

If the exact site and its neighboring amino acids were conserved in any of the orthologs from the three pre-WGD species, or in the other ohnolog (for WGD proteins only), then we inferred that this site was also present in their last common ancestor that was living just before the WGD, ~ 100 MYA (see Figure 3.3). It has been shown that the neighboring three amino acids to the left and the three to the right of the exact phosphorylation site are more conserved than the average (Gnad, Ren et al. 2007). Therefore, taking into account this information, we generated eight different datasets (we call them ancestral) of ancestral p-sites, by arbitrarily allowing a variation of 2, 3, 4 and 6 amino acids in the designated vicinity of the p-site, for both 6eNU and 6eU, as long as the p-site was not mutated.

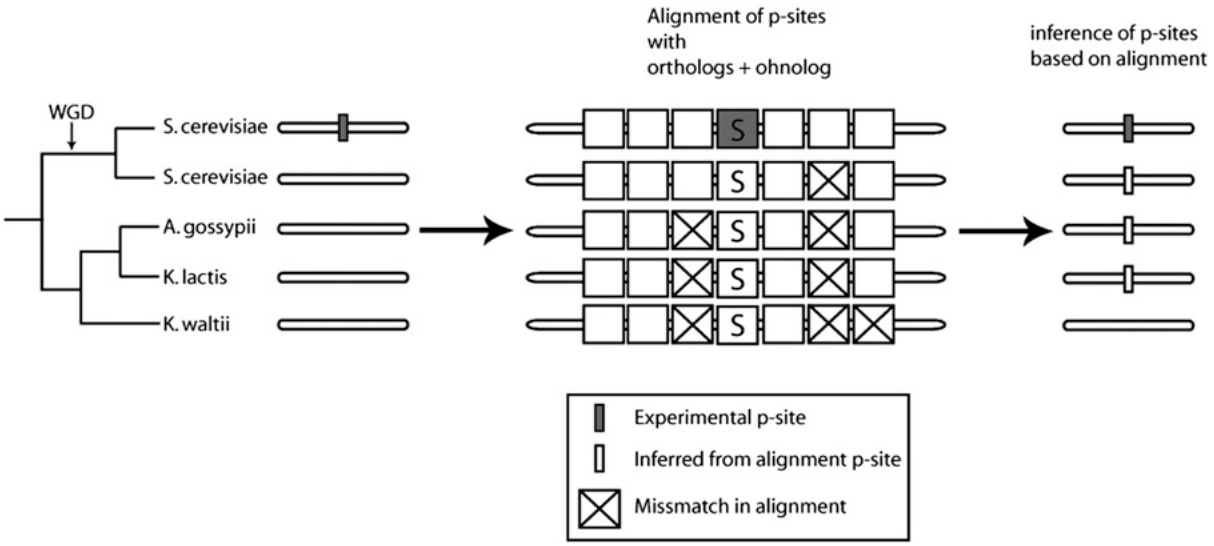


Figure 3.3: Inference of p-sites in the pre-WGD orthologous ancestral protein. The inference is based on alignment with pre-WGD orthologs (and ohnolog, whenever appropriate). Here, the threshold for inference is no more than two amino acid mismatches in the window of six amino acids, surrounding the p-site.

By using the Yeast Genome Order Browser, we also identified the orthologs of *S. cerevisiae* WGD and RSS-genes in each of the available three genomes of species (*S. bayanus, C. glabrata and K. polysporus*) that diverged after the WGD event. For each of the ancestral (pre-WGD) genes, we identified in how many of the four post-WGD species, the duplicates were retained. Products of genes that survived as duplicates in the genomes of at least three of the four post-WGD yeast species are designated 'retained-in-majority', whereas products of genes that have returned to single-copy status in the genomes of at least three of those four post-WGD species are designated 'lost-in-majority'. Thus, we compared the ancestral pre-WGD orthologs of the 'retained-in-majority' bin versus the 'lost-in-majority' bin, for each of the eight ancestral p-site datasets

3.5 Supporting Information

3.5.1 Two phosphorylation datasets:

Experiment	Albuquerque et al.	Bodenmiller et al.	Chi et al.	Gruhler et al.	Reinders et al.
For peptides that exactly match more than one proteins					
Total p-sites	4185	6036	1442	1036	78
P-sites verified by other/s	2035	2827	807	759	17
%verified	48.6	46.8	56.0	73.3	21.8
For peptides that exactly match only one protein					
Total p-sites	3365	4908	727	677	78
P-sites verified by other/s	1692	2250	400	498	17
%verified	50.3	45.8	55.0	73.6	21.8

SI.Table 3.1: This table summarizes how many p-sites each experiment contributed and how many of them were also verified by at least one other experiment. The dataset of Bodenmiller et al., (2008) also contained the dataset of Li et al., (2007). Therefore, we could calculate the contribution of the Bodenmiller dataset on its own, but not of the Li dataset.

We used 6 experimental data sets(Gruhler, Olsen et al. 2005; Chi, Huttenhower et al. 2007; Li, Gerber et al. 2007; Reinders, Wagner et al. 2007; Albuquerque, Smolka et al. 2008; Bodenmiller, Campbell et al. 2008), applying to every one of them the filter proposed in that specific study, for identifying the exact location of phosphorylation sites. All the above experiments rely on affinity-based methods (IMAC) for phosphopeptide isolation and their results are estimated to be up to 93% reproducible(Bodenmiller, Mueller et al. 2007). The identified phosphopeptides could match (exactly) either one or more than one ORFs. Accordingly, we generated two phosphorylation datasets: One that contains phosphopeptides, that have a unique match (designated as 6eU), and one that contains phosphopeptides that may exactly match one or more ORFs (designated as 6eNU). The reason is that this data treatment may have an effect on analyses of gene duplication. Transposable elements were removed from the two datasets. Overall, the 6eNU dataset contained 11,327 phosphorylation sites (hereafter referred to as p-sites) in 2453 proteins (now called phosphoproteins). The 6eU dataset overall contained 8513 p-sites in 2289 phosphoproteins.

3.5.2 GO-slim enhancement analysis of the phosphorylation datasets

GO-slim enhancement analysis of the two datasets (6eNU and 6eU) was performed with the BINGO cytoscape plug-in (Maere, Heymans et al. 2005). A hypergeometric test was used, at the p=0.05 significance level, after applying the Benjamini-Hochberg multiple-testing correction. Results (for 6eNU and 6eU) are shown in SI.Table 3.2& 3.3 below.

6eNU		
GO-ID	Description	Corrected p-value
5634	nucleus	7.08e-23
6996	organelle organization and biogenesis	9.03e-23
134	site of polarized growth	4.87e-22
5933	cellular bud	4.98e-20
5938	cell cortex	4.00e-17
5856	cytoskeleton	5.76e-15
6350	transcription	8.28e-15
5840	ribosome	9.75e-15
7165	signal transduction	5.87e-13
50222	protein kinase activity	2.18e-12
5198	structural molecule activity	4.69e-12
7010	cytoskeleton organization and biogenesis	1.11e-11
9653	anatomical structure morphogenesis	1.11e-11
30528	transcription regulator activity	2.62e-11
7049	cell cycle	5.78e-11
16192	vesicle-mediated transport	2.22e-10
6950	response to stress	5.36e-10
5737	cytoplasm	6.29e-10
7114	cell budding	1.27e-09
5730	nucleolus	2.74e-09
30234	enzyme regulator activity	7.22e-09
42254	ribosome biogenesis	7.72e-08
45182	translation regulator activity	6.45e-06
16070	RNA metabolic process	9.76e-06
3677	DNA binding	9.81e-06
6416	translation	4.95e-05
5694	chromosome	5.16e-05
16288	cytokinesis	5.37e-05
16044	membrane organization and biogenesis	5.79e-05
6810	transport	2.46e-04
5815	microtubule organizing center	4.16e-04
6997	nuclear organization and biogenesis	6.78e-04
7124	pseudohyphal growth	8.62e-04
746	conjugation	9.13e-04
5773	vacuole	9.96e-04
5886	plasma membrane	1.67e-03
5515	protein binding	1.73e-03
30163	protein catabolic process	3.42e-03
4871	signal transducer activity	4.06e-03
4386	helicase activity	8.58e-03
5624	membrane fraction	4.74e-02
6464	protein modification process	4.88e-02

SI.Table 3.2: GO-slim enhancement analysis for dataset 6eNU.

6eU		
GO-ID	Description	Corrected p-value
5634	nucleus	8.43e-32
134	site of polarized growth	1.66e-24
6996	organelle organization and biogenesis	3.21e-24
5933	cellular bud	1.47e-22
6350	transcription	4.41e-20
5938	cell cortex	9.85e-18
5856	cytoskeleton	3.56e-16
7049	cell cycle	7.74e-15
50222	protein kinase activity	1.55e-14
7010	cytoskeleton organization and biogenesis	1.55e-14
7165	signal transduction	1.55e-14
30528	transcription regulator activity	1.55e-14
9653	anatomical structure morphogenesis	1.55e-13
16192	vesicle-mediated transport	2.45e-11
30234	enzyme regulator activity	3.86e-11
7114	cell budding	4.68e-11
6950	response to stress	4.82e-11
16070	RNA metabolic process	5.23e-09
5730	nucleolus	2.77e-07
3677	DNA binding	1.10e-06
42254	ribosome biogenesis	2.44e-06
16044	membrane organization and biogenesis	6.62e-06
5737	cytoplasm	9.42e-06
16288	cytokinesis	9.42e-06
5694	chromosome	1.79e-05
6810	transport	5.21e-05
5815	microtubule organizing center	8.41e-05
746	conjugation	1.95e-04
5515	protein binding	3.59e-04
7124	pseudohyphal growth	4.16e-04
6997	nuclear organization and biogenesis	5.63e-04
5886	plasma membrane	8.26e-04
30163	protein catabolic process	8.36e-04
5773	vacuole	8.63e-04
6464	protein modification process	1.60e-03
45182	translation regulator activity	2.05e-03
4871	signal transducer activity	4.28e-03
4386	helicase activity	5.23e-03

SI.Table 3.3: GO-slim enhancement analysis for dataset 6eU.

3.5.3 WGD and RSS duplication datasets for *S. cerevisiae*

We analysed genes of *S. cerevisiae*, a species that underwent a WGD 100 mya. Genes were separated into those that were retained in duplicate following the WGD until today (designated as WGD-genes) and those that underwent the duplication but later lost the duplicate (designated as Returned to Single Status: RSS-genes). A duplicate of a gene that resulted from the WGD event is called ohnolog. The

assignment of orthologs and ohnologs was based on syntenic information(Byrne and Wolfe 2005), using the *S. cerevisiae* genome (Goffeau, Barrell et al. 1996) as well as genomes that did not undergo the WGD. SI.Table 3.4 summarizes the statistics.

	WGD	RSS
total proteins	1096	4002
phosphoproteins in 6eNU	643	1733
phosphoproteins in 6eU	530	1704
p-sites in 6eNU	3075	5992
p-sites in 6eU	2458	5933

SI.Table 3.4: Statistics for WGDs and RSSs.

3.5.4 *In-silico* prediction of p-sites in yeast proteins

We have attempted a prospective analysis by using the NetPhosYeast software (Ingrell, Miller et al. 2007) to predict protein phosphorylation sites for the entire *Saccharomyces cerevisiae* proteome. The NetPhosYeast software is specifically tailored to *S. cerevisiae* and has been shown to outperform all other predictors. We ran the predictions with two different stringent cut-offs (0.75 and 0.85) and determined that WGD-phosphoproteins have more p-sites on average than RSS-phosphoproteins and that this difference is statistically significant (data not shown).

3.5.5 The selection, quality and evolution of the datasets do not affect our conclusions

A jackknife analysis considering all 6 experimental datasets showed that no matter which dataset is excluded from the analysis, WGD-phosphoproteins have on average more p-sites than RSS-phosphoproteins. Therefore, our conclusions are not sensitive to any of the datasets, even if any of them would be of questionable quality. In addition, we excluded the two most recent datasets (Albuquerque, Smolka et al.

2008; Bodenmiller, Campbell et al. 2008) and observed that not only is the trend observed, but also, as more datasets become available, the difference in p-sites among WGD and RSS-phosphoproteins increases, from the 22-25% level in the year 2007, at the 33-38% level in the year 2008. A further re-assurance that our conclusions are not affected by the incompleteness of the experimental data is the confirmation from the analysis of the Netphosyeast predicted dataset.

3.5.6 Analysis for testing whether the trend is general and not sensitive to certain gene groups (GO_Slim categories)

First, we wanted to test, for each of the 2 phosphorylation datasets (6eNU & 6eU), whether the observed trend between WGD- and RSS-phosphoproteins (at the genome scale) is also found within each of the various GO-Slim categories. For reasons of statistical stringency, we used only those GO-Slim categories that have at least 30 phosphoproteins in each of the WGD and RSS groups. A Wilcoxon test was used, for each GO-Slim category, at the 0.05 significance level. Next, we applied Benjamini-Hochberg multiple testing correction in the obtained p-values. Next, we wanted to rule out the possibility that the observed trend appears as general, due to a strong effect within only one category. For this reason, we implemented a Jackknife approach, where each one of the GO-slim categories (without implementing a threshold of 30 phosphoproteins, as applied above) was removed from the dataset and the Wilcoxon test was run, again at the 0.05 significance level, followed by the Benjamini-Hochberg multiple testing correction. The same was applied for Pfam domains that are grouped together, based on common evolutionary ancestry, based on the Superfamily database (Wilson, Pethica et al. 2009). Again, the results were not sensitive to the removal of any superfamily-domain (results not shown). We also repeated the Jackknife approach by simultaneously removing all transcription factors (TFs), kinases, phosphatases and cyclins of *S. cerevisiae* simultaneously. For assignment of these regulatory molecules we used the DBD database (Wilson, Charoensawan et al. 2008) and the website (http://kinase.com) (Manning, Plowman

et al. 2002). The same analysis was performed by removing the ribosomal proteins (as identified from the SGD database; http://www.yeastgenome.org). Given that the subunits of protein complexes are often phosphorylated, it is possible that the preferential retention of WGD pairs encoding highly phosphorylated proteins is a by-product of the need to maintain a stoichiometric balance between subunits (Papp, Pal et al. 2003; Mintseris and Weng 2005). Therefore, we repeated (for 6eNU and 6eU) the statistical analysis after removing all proteins that are members of complexes (according to the latest curated dataset (Pu, Wong et al. 2009) for *S. cerevisiae*).

3.5.7 Controlling for biases

3.5.7.1 Protein abundance

By using a published large-scale dataset on protein abundances that were expressed during exponential-phase growth (Ghaemmaghami, Huh et al. 2003), we performed an analysis to exclude the possibility that our conclusions are biased by protein abundance. We identified 13 GO_slim individual categories for which the WGD-group has on average higher protein abundance than the RSS-group.

GO_slim ID	Description
GO:0005198	structural molecule activity
GO:0006412	translation
GO:0005840	ribosome
GO:0042254	ribosome biogenesis
GO:0006996	organelle organization
GO:0005737	cytoplasm
GO:0006950	response to stress
GO:0005975	carbohydrate metabolic process
GO:0006725	cellular aromatic compound metabolic process
GO:0006457	protein folding
GO:0006519	cellular amino acid and derivative metabolic process
GO:0016874	ligase activity
GO:0003723	RNA binding

Next, we excluded them from the two datasets (6eNU and 6eU) and repeated the analyses for difference in phosphorylation among WGD and RSS groups. We checked that the new WGD-groups indeed did not have higher protein abundance and then we verified for both datasets that the trend is still observed and statistically significant.

3.5.7.2 Coverage in experiments

It may be that WGD-phosphoproteins are present in more conditions than RSS-phosphoproteins, although there could be no significant difference among them. If that were the case, then the more widely present phosphoproteins could be accumulating more p-sites in the various experiments. To control for this, we generated one dataset by using only two of the large-experiment datasets (Gruhler, Olsen et al. 2005; Chi, Huttenhower et al. 2007; Li, Gerber et al. 2007; Albuquerque, Smolka et al. 2008; Bodenmiller, Campbell et al. 2008). We analysed only the 742 phosphoproteins (196 WGD+ 546 RSS) that have at least one phosphopeptide found in both datasets; this phosphopeptide must have an exact match with only one ORF. Next, we added (for that protein) the p-sites that have been detected in either of the two experiments. Another strong re-assurance that our conclusions are not an artifact of experimental biases is their confirmation from the analysis of the Netphosyeast-predicted phosphorylation dataset.

3.5.7.3 Essentiality of genes

It has been demonstrated that in yeast, essential genes are older than nonessential genes and form an exponential core in the yeast protein interaction network (Pereira-Leal, Audit et al. 2005). To rule out the possibility that essentiality is actually the primary factor for the observed trend, we excluded from our dataset all those essential genes that have been confirmed both by two large scale knock-out

experiments (Giaever, Chu et al. 2002; Steinmetz, Scharfe et al. 2002; Pache, Babu et al. 2009).

3.5.7.4 Protein interaction network centrality

We also excluded from our datasets all the hubs of the yeast protein interaction network. We used the high confidence dataset of (Batada, Reguly et al. 2006) and used two different cutoffs to define hubs, with connectivity k<21 and k<10.

3.5.8 Inferring the phosphorylation sites of the pre-WG duplication ancestor

Based on orthology-paralogy identified by previous workers (Byrne and Wolfe 2005), we aligned each *S. cerevisiae* protein (and ohnolog, whenever appropriate) with its orthologs from the 3 pre-duplication species (Dietrich, Voegeli et al. 2004; Dujon, Sherman et al. 2004; Kellis, Birren et al. 2004) (*Ashbya [Eremothecium] gossypii*, *Kluyveromyces lactis*, *Kluyveromyces waltii*), using the t-coffee software (default parameters) (Notredame, Higgins et al. 2000). If the exact site and its neighbouring amino acids were conserved in any of the orthologs from the 3 pre-WGD species, or in the other ohnolog (for WGD proteins only), then we inferred that this site was also present in their last common ancestor that was living just before the WGD, ca. 100 MYA. It has been shown that the neighbouring three amino acids to the left and the three to the right of the exact phosphorylation site are more conserved than the average (Gnad, Ren et al. 2007). Therefore, taking into account this information, we generated 8 different datasets (we call them ancestral) of ancestral p-sites, by arbitrarily allowing a variation of 2, 3, 4 and 6 amino acids in the designated vicinity of the p-site, for both 6eNU and 6eU, as long as the p-site was not mutated. Next, we tested for each one of those eight ancestral datasets, if the pre-WGD ancestral

ortholog that later gave rise to two WGD-genes had, on average, more ancestral psites than the pre-WGD ancestral ortholog that later duplicated but lost the second copy (RSS gene). To control for the fact that we used two ohnologs but only one RSS-protein, for ancestral p-site inference, we repeated the ancestral inference, this time using only one (that with the most p-sites) of the two ohnologs. Four different ancestral datasets were generated, this time using only the 6eU dataset, for variation of 2, 3, 4 and 6 amino acids. Then, a comparison of WGD and RSS ancestors was performed, as in the previous test. By applying the same methodology explained above (for both 6eU and 6eNU, and for variation 2, 3, 4 and 6 amino acids), we also projected the ancestral p-sites in the other ohnolog, if the exact position was not an experimentally determined p-site but, nevertheless, the criterion for variation was met. In this way, we generated 8 different datasets again, for the ohnologs and determined which ancestral p-sites were reciprocally lost. Thus, we tried to minimize problems from incompleteness of phosphorylation data. Then, we tested if the ancestors of sub-functionalised pairs had on average more p-sites than the other ancestors. This evolutionary analysis also constitutes an additional control, since not all experimentally identified p-sites in yeast need to affect the function of the protein. We used p-sites that have been conserved over 100 million years and most probably have been important for the function of the protein, otherwise they would be wiped out by mutations.

3.5.9 The number of ancestral p-sites affects the retention of WG duplicates in independent post- WGD lineages

By using the Yeast Genome Order Browser (Byrne and Wolfe 2005), we identified the orthologs of *S. cerevisiae* WGD and RSS-genes in each of the available 3 genomes of species (*S.bayanus*, *C.glabrata* and *K.polysporus*) that diverged after the WGD event. For each of the ancestral (pre-WGD) genes, we identified in how many of the 4 post-WGD species, the duplicates were retained. Products of genes that survived as duplicates in the genomes of at least 3 of the 4 post-WGD yeast species

are designated ‘retained-in-majority’, whereas products of genes that have returned to single copy status in the genomes of at least 3 of those 4 post-WGD species are designated ‘lost-in-majority’. Thus, we compared the ancestral pre-WGD orthologs of the ‘retained-in-majority’.

3.5.10 Secondary structure and intrinsic disorder prediction

We used the SABLE software (Wagner, Adamczak et al. 2005) (http://sable.cchmc.org/), as was done in a previous analysis (Gnad, Ren et al. 2007), to identify three categories of secondary structure; alpha-helices, beta-strands and loops. For identifying intrinsically disordered regions, we used the VSL2B software (Peng, Radivojac et al. 2006) that was ranked as first in CASP7. It has been proven previously (Iakoucheva, Radivojac et al. 2004) that phosphorylation sites are more often found in loops and flexible, unstructured regions.

3.5.11 Ubiquitination and protein half-lives

The analysis on ubiquitination was based on a large-scale experiment (Peng, Schwartz et al. 2003) in *S. cerevisiae*. This is only one experiment and does not identify the exact positions of ubiquitination. The values of protein half-lives were taken from a published large-scale experiment (Belle, Tanay et al. 2006) in *S. cerevisiae*.

3.5.12 Small-scale gene duplications versus singletons in *S. cerevisiae*

	SSDs	Singletons
total proteins	2260	2439
phosphoproteins in 6eNU	986	824
phosphoproteins in 6eU	944	815
p-sites in 6eNU	3665	2510
p-sites in 6eU	3556	2496

SI.Table 3.5: Statistics for SSDs and singletons.

From the 5795 yeast protein coding genes (excluding transposable elements), we first removed the 1096 WGD genes. Next, we defined 2439 *S. cerevisiae* genes as singletons, based on the fact that their proteins had no Blastp hit against any other *S. cerevisiae* proteins, at a cut-off level of $1e^{-3}$. As Small-Scale gene Duplicates (SSDs), we identified 2260 genes that belong to neither the WGD group, nor the Singletons group. SI.Table 3.5, below, summarizes the statistics.

3.5.13 Small-scale gene duplications versus singletons in *Schizosaccharomyces pombe*

We defined 2240 *Schizosaccharomyces pombe* genes as singletons, based on the fact that their proteins had no Blastp hit against any other *Sz. pombe* proteins at a cut-off level of $1e^{-3}$. Since this organism is not known to have undergone WGD, we assign the other 2782 protein-encoding genes to the SSD group. We used a large-scale phosphoproteomic dataset (Wilson-Grady, Villen et al. 2008) to identify phosphoproteins and their exact p-sites, using the proposed cut-off (A score>19). The experiment was performed twice, with two different technologies, titanium dioxide (TiO_2) and IMAC. From the union of the two technologies and after filtering for exact p-sites, 1711 non-redundant exact p-sites were assigned to 941 proteins. A higher fraction (22.5%) of SSD proteins were phosphorylated compared to singletons (14%), and this difference was statistically significant ($p<2e^{-14}$, chi-squared). Nevertheless, when we tried to investigate, whether there was a difference in the absolute number of p-sites in the phosphoproteins among the SSD and singleton

groups, we found none (Wilcoxon, p=0.23). We attribute this to the paucity of the *Sz. pombe* data, the dataset is ~3.5 times smaller than that for *S. cerevisiae*. For this reason, we performed 1000 simulations, where we randomly reduced the number of p-sites for *S. cerevisiae* to the level observed in *Sz. Pombe* (1711 exact p-sites) and found that in 496/1000 simulations, the Wilcoxon test would fail to show the difference between SSD and Singleton phosphoproteins. Nevertheless, for the same simulated datasets, the SSD group always had a higher fraction of phosphoproteins, compared to the singleton group, and this difference was always supported statistically by the chi-square test. In conclusion, even small datasets can reveal a difference in the fraction of phosphoproteins among the SSD and singleton groups. Nevertheless, the Wilcoxon tests need more data to verify that SSD-phosphoproteins have more p-sites (on average) than singleton proteins.

3.6 Author contributions

Y.H. helped prepare the dataset and conducted part of the bioinformatics analysis under the supervision of G.A. and Y.V.D.P.

4 Preferential substitutions of phosphorylation sites in eukaryotes

He, Y., Amoutzias, G., Van de Peer,Y.

Manuscript under preparation

Abstract

Post-translationally modified amino acids and in particular phosphorylated serines (pS) might have evolved differently from their non-modified counterparts. Previous studies have demonstrated that mutation of Serine or Threonine (S/T) to Aspartic acid or Glutamic acid (D/E) has a similar effect to these two amino acids (S/T) being phosphorylated. By employing the yeast phosphoproteome, this study investigates the pattern of pS substitutions through pairwise alignments between *S. cerevisiae* genes and any of their orthologs in another six fungal species of gradient evolutionary distances. Concerns that may compromise the conclusions of this study have been rigorously addressed, such as noisy phosphorylation data (e.g. technical false positive or low-stoichiometry off-target phosphorylation sites) as well as low-quality of alignment between homologous sites. Interestingly, Serines that are targeted by kinases for phosphorylation tend to be substituted to Aspartic acid and Glutamic acid more frequently than Serines not targeted by kinases. In addition, the same category of Serines (targeted by kinases for phosphorylation) tends to be substituted to Alanine less frequently than Serines not targeted for phosphorylation. This intriguing pattern of substitutions reveals the constraints that are enforced on the evolution of phosphorylated proteins and could act as an additional type of genomic information for the prediction of phosphorylated sites.

Similarly, the substitution patterns of pS have been investigated in animals and plants. However, there was no strong evidence for the pattern that was originally observed in fungi. This could be attributed to the fact that the current compendiums of both human and *A.thaliana* phosphoproteomes are far from being complete. Therefore, the negative (non-phosphorylated) datasets are most probably “contaminated” with undetected phosphorylation sites and therefore cannot reveal the “true” properties and substitution patterns of the phosphorylated serines.

4.1 Introduction

Protein phosphorylation is the addition of a phosphate (PO_4) group to the -OH group of serine (S), threonine (T), or tyrosine (Y) of a protein, changing -OH to -PO_4^{2-}. This addition is mediated by kinases whereas the removal of the phosphate is mediated by phosphatases. Reversible phosphorylation of proteins is an important regulatory mechanism that occurs in both prokaryotic and (more prevalently in) eukaryotic organisms (Chang and Stewart 1998; Garske, Peters et al. 2011). The attachment of the phosphate group adds a negative charge to the protein, that may lead to an allosteric conformational change in the structure of the protein, with its function being either switched 'on' or 'off'. Signal transduction is mediated by phosphorylation-prompting conformational changes in the target protein and thus, via a cascade of changes, a signal reaches the nucleus. Phosphorylation has a prominent role in signaling (Li, Wang et al. 2000; Mustilli, Merlot et al. 2002; Yoshida, Hobo et al. 2002), cell differentiation, motility and proliferation, among many other biological processes (Roskoski 2005). Previous studies have shown that mutating a Serine or Threonine (S/T) to Aspartic acid or Glutamic acid (D/E) has a similar effect to these amino acids (S/T) being phosphorylated (Furihata, Maruyama et al. 2006). That is, the substitution of a S/T to D/E is equivalent to having this residue always phosphorylated. On the contrary, a S/T to A mutation has the same effect as dephosphorylation (Furihata, Maruyama et al. 2006). This effect is attributed to the alanine resembling a nonphosphorylated serine/threonine. The difference in the chemical structures between alanine (A) and serine (S) is the missing -OH group. Alanine (compared to threonine) lacks both the –OH and $–CH_3$ group (Figure 4.1 above). Aspartic acid (D)/glutamic acid (E), on the other hand, resembles to phosphorylated S/T both in chemical structure and charge (see Figure 4.1 below (Anthis 2009)).

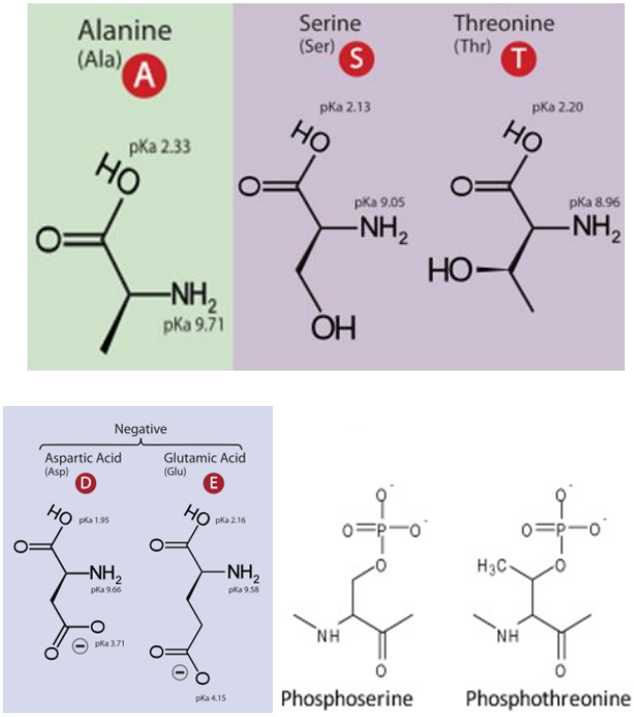


Figure 4.1: **Above**: the chemical structures of alanine (A), serine (S) and threonine (T) are shown, respectively; **below**: the chemical structures of aspartic acid (D), glutamic acid (E) and phosphorylated serine, threonine are shown (taken from (Anthis 2009)).

Reversible phosphorylation of S/T constitutes a regulatory event, but a substitution of S/T to D/E essentially compromises this regulatory event. Among diverging organisms, different types of molecular networks change (rewire) at different rates, but with a decreased rate of rewiring over certain time of divergence because of saturation in potential substitutions (Shou, Bhardwaj et al. 2011). When such a substitution occurs, an edge (one phosphorylation) between two nodes (kinase and substrate) will be lost and the structure of the underlying regulatory network will be changed. The same applies for the phosphatase-target network.

This study employs a curated compendium of twelve publicly available high-throughput MS-based phosphoproteome data from the model organism Saccharomyces cerevisiae (Gruhler, Olsen et al. 2005; Chi, Huttenhower et al. 2007; Li, Gerber et al. 2007; Albuquerque, Smolka et al. 2008; Bodenmiller, Campbell et al. 2008; Beltrao, Trinidad et al. 2009; Gnad, de Godoy et al. 2009; Holt, Tuch et al. 2009; Huber, Bodenmiller et al. 2009; Soufi, Kelstrup et al. 2009; Stark, Su et al.

2010) to study the substitution-patterns of phosphorylation-sites. In particular, we investigate whether the substitution of phosphorylated serines is under selective pressure, compared to the substitution of non-phosphorylated serines and if so, whether there is a bias towards specific amino acids with similar properties. Phosphorylation data from human and the model plant A. thaliana are also investigated to conclude whether any preferential substitution pattern is limited within the fungal lineage or whether it is common among the major eukaryotic lineages.

4.2 Results and Discussion

4.2.1 Substitution patterns in Fungi

4.2.1.1 Summary for the full pS and npS substitution

The high quality dataset of p-sites that is assembled in Chapter 2 where very stringent criteria have been applied to filter out both technical false-positives and low-stoichiometry off-target phosphorylations is used in this study (see Materials and Methods, section 4.4.3). The issues of false negatives (uncovered p-sites) is also addressed by applying a filter to weed-out serines with high Netphosyeast prediction scores (>0.7) from non-phosphorylation sites, thus creating a second negative dataset designated npS2, (see section 4.4.3).

Previous analyses show that most p-sites are found in disordered regions (Landry, Levy et al. 2009). We also observed that 91% of STY p-sites are found in disordered regions, compared to 54% of non-phosphorylated STY sites (Chapter 2). Intrinsically disordered regions are fast evolving. Therefore, we repeat our analysis on both the “full” dataset and the “disorder” (disordered regions only) to control for different

evolutionary rates of serines. In addition, the phosphopeptides undergo an enrichment step, where acidic peptides are preferentially kept. It is conceivable that p-sites may incorrectly align with other acidic amino acids in the neighbouring positions of the exact substituting sites. Therefore, we take into account for the quality of the alignment (3 categories: no_filter, blast_filter, aln_filter) in this study (see Materials and Methods, section 4.4.5). General statistics of pS and npS substitutions is summarized in Supporting Information. When no filter ("no_filter") is applied to sequence conservation, the numbers of phosphorylated serines (pS) that are used to study substitutions are over 7000. When "blast_filter" is applied, the numbers of serines used to extract substitutions for different dataset are comparable to the "no_filter" dataset (a few hundreds less than "no_filter"). However, when the more stringent strategy "aln_filter" is applied, the total number (~2000) of pS used in different datasets dropped significantly. Therefore, the results from "blast_filter" are displayed in the following sections where the issue of the quality of alignment is addressed and major conclusions of this study are mainly based on those runs with "blast_filter". Results from "no_filter" and "aln_filter" are available in the Supporting Information.

The percentages of S to D/E/A substitutions for pS, npS1 and npS2 for the "blast_filter" and "full" dataset range from 0.3% - 6.7% (see Table 4.1). The observed differences in S to D/E/A substitutions between the pS and npS groups are also tested for significance in the following sections. Notably, the percentage of serine to alanine (S to A) substitution is larger in the negative datasets, compared to the positive datasets (over 1%, Table 4.1), which is also interesting because alanine (A) mimics the non-phosphorylation state.

Dataset	Alignment	AA	Sbay	Cgla	Scas	Kpol	Zrou	Klac
pS	Blast	D	1.20	4.30	3.75	4.25	3.99	4.30
npS1			0.79	3.34	2.81	3.02	2.76	3.27
npS2			0.76	3.25	2.74	2.96	2.64	3.20
pS		E	0.49	4.03	4.10	3.99	4.32	4.59
npS1			0.26	3.25	2.80	3.10	2.91	3.41
npS2			0.25	3.18	2.72	3.05	2.81	3.41
pS		A	1.74	4.81	4.24	4.14	4.92	4.87
npS1			2.58	6.43	5.79	5.16	6.72	6.50
npS2			2.61	6.68	5.97	5.34	6.98	6.74

Table 4.1: The percentages (%) of serines substitutions (S to D, E, A) in the run **blast_filter**, pS, npS1, npS2, **full**.

4.2.1.2 Difference between pS and npS

The overall percentages of serine substitutions (only to D, E and A) for pS, npS (npS1 and npS2) and reS (resampling from npS) are displayed in the following sections, with control for the quality of alignment ("blast_filter"), and for S to D, E and A substitutions, respectively.

In the "ori" resampling, a bias exists within resampling, which is indicated by deviation of average % of npS substitution from the reS distribution curve (see SI.Figure 4.8 in Supporting Information). Therefore, despite the fact that a trend for a preferential substitution is observed in most of runs with "ori" resamplings, it might not be a good way to evaluate the trend for preferential substitutions of pS, where the exact number of p-site for each p-protein is accounted for resampling. Nevertheless, the results for the "ori" resampling are presented (SI.Figure 4.8) to give complementary opinions.

To summarize, the datasets that are used to reveal the pattern of pS substitutions in this study are the runs with "blast_filter", "re_1000" resampling, which according to us, suffer less from false positives (HQ pS) and false negatives (npS2) of p-sites and certain biases in low quality of alignment (blast_filter) and resampling (re_1000).

4.2.1.2.1 Preferential substitution of pS to D and E

The overall percentages of pS to D substitutions and the significance for difference against the negative dataset are shown in Figure 4.2 (negative dataset: npS1) and Figure 4.3 (negative dataset: npS2). For this analysis the “blast_filter” and “re_1000” resampling criteria were used. There are obviously significant differences in S to D substitutions between pS and npS (both for npS1 and npS2, indicated by the * in each figure).

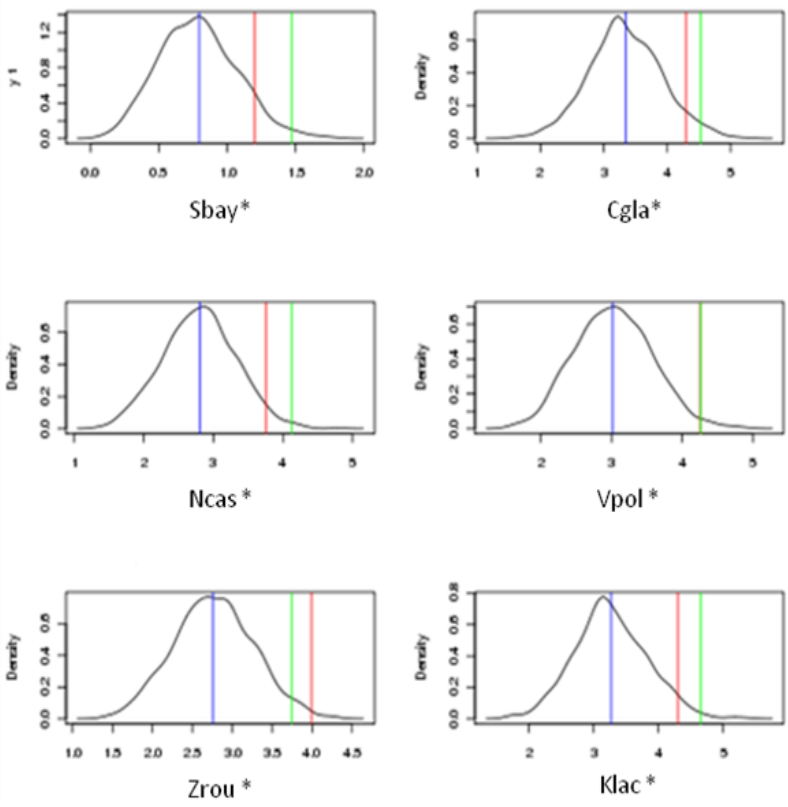


Figure 4.2: The overall percentages of serine substitutions (x axis) is plotted against its density distribution (y axis). The percentages for substitution of pS (red line), npS (blue line), reS (distribution) and the 75% tail of reS (green line) are displayed for each of the six fungal species (from the closest post-WGD species *Saccharomyces bayanus* to less close post-WGD species (in a descending order of closeness) *Candida glabrata, Naumovia castellii,* and *Vanderwaltozyma polysporus*, and pre-WGD species *Zygosaccharomyces rouxii* and *Kluyveromyces lactis*), respectively. The star (*****) next to species names represents a significant p-value in Mann-Whitney U test (see Materials and Methods).The figure is for the run: **S to D substitution**, **blast_filter**, **npS1**, **full**, **re_1000 resampling**. Probabilities (see Materials and Methods) for such a preferential S to D substitution of these 6 species are 0.947, 0.955, 0.963, 0.981, 0.993 and 0.955. In the case of *V.pol,* the % of pS substitution is equal to the 75% tail of reS (red line overlaps with green line).

The trend for a preferential substitution of pS in the run with npS2 is stronger than the one with npS1 (see Figure 4.2 and Figure 4.3). npS2 is a subset of npS1 but with

false negatives of p-sites being removed by the netphosyeast prediction algorithm. In this study, both the results from npS1 and npS2 are considered when conclusions are drawn.

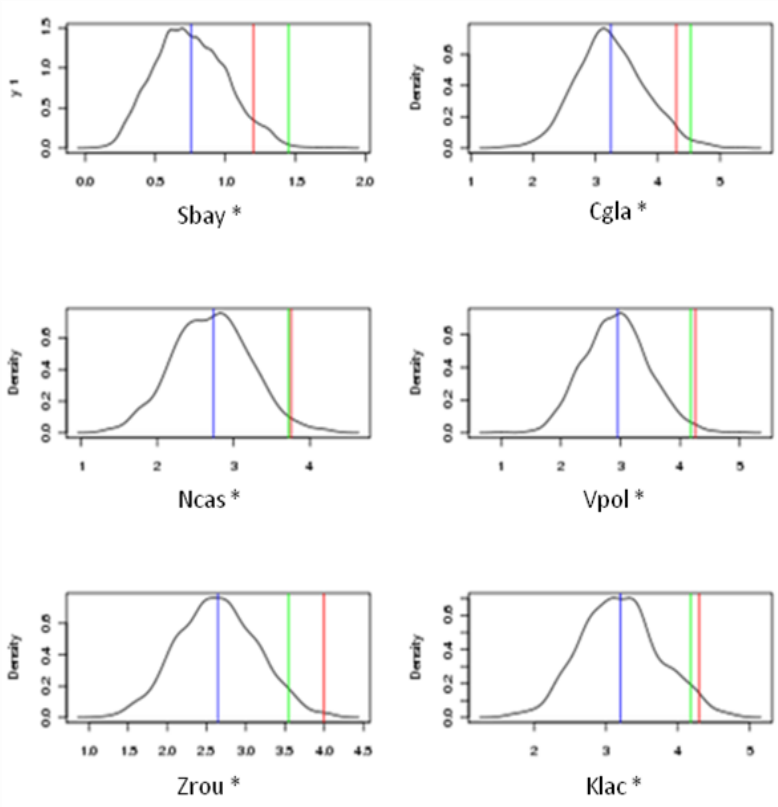


Figure 4.3: Similarly to Figure 4.2, this figure is for the run: **S to D substitution**, **blast_filter**, **npS2**, **full**, **re_1000 resampling**. Probabilities for such a preferential S to D substitution of these 6 species are 0.962, 0.975, 0.977, 0.99, 0.993 and 0.966.

The trend for a preferential substitution from pS to E is also strong (see Supporting Information, SI.Figure 4.2 and SI.Figure 4.3). We have not found such patterns of preferential substitution of pS to other amino acids (other than D, E and A) in this study (data not shown). Significant differences (measured by p-values in Mann-Whitney U test, see Methods and Materials) in the % of substitutions between pS and npS are observed for almost all these runs for substitution to D/E (with only one exception, in the run with “aln_filter” and disorder regions only, data not shown). Among different filters on the conservation of sequence alignment, the trend for a preferential substitution from ““aln_filter” is less strong (see SI.Figure 4.6 and

SI.Figure 4.7), which might be due to the significant decrease in the number of substitutions in datasets of distant species (SI.Table 4.2). Results from “no_filter” (SI.Figure 4.4 and SI.Figure 4.5) and disordered regions “disorder” also support the preferential substitution of pS to both D and E (SI.Figure 4.9).

4.2.1.2.2 “Disfavored” substitution of pS to A

Alanine (A) is a non-phosphorylation mimic, thus resembling the non-phosphorylation status of serines and threonines. As shown early in this study, pS tend to be more frequently substituted to D/E compared to the non-phosphorylated serines. A less frequent (or “disfavored”) substitution of pS to A is also observed in the dataset with “blast_filter”, npS2, full and “re_1000” (Figure 4.4).

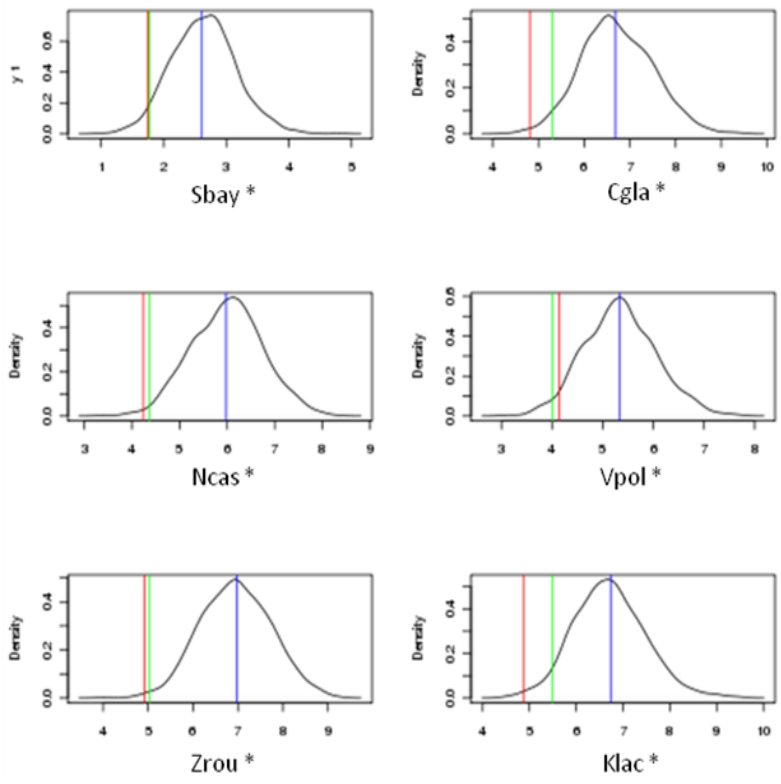


Figure 4.4: Similarly to Figure 4.2, but the green line is for the lower 25% tail of reS distribution. The figure is for the run: **S to A substitution**, **blast_filter**, **npS2**, **full**, **re_1000 resampling**. Probabilities for such a preferential S to A substitution of these 6 species are 0.967, 0.995, 0.992, 0.963, 0.995 and 0.992.

4.2.2 Substitution patterns in vertebrates and plants

Similar approaches were applied to studies with vertebrates and plants. The overall percentage (%) for substitutions are shown in Figure 4.4 and Figure 4.5, for vertebrates and plants (blast_filter, re_1000 resampling, full length, S to D substitutions), respectively. We have not observed any significant differences in the % of serine substitutions of pS and npS nor a high probability for preferential substitutions of pS using human phosphorylation dataset (Figure 4.4). The result is not surprising because the current collection of pS dataset in human is far from being complete; therefore, there are many undetected p-sites contaminating the npS dataset for human.

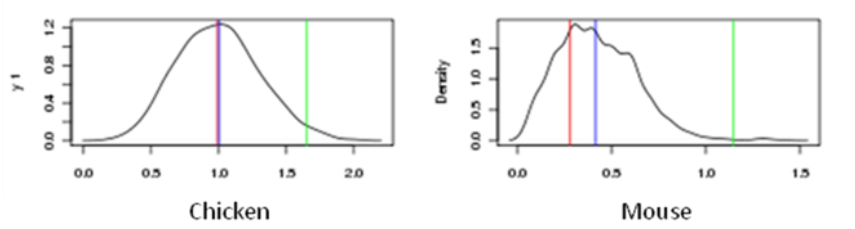


Figure 4.5: Similarly to Figure 4.2, this figure is for the run: **S to D substitution, blast_filter**, **full**, **re_1000 resampling**. Probabilities for such a preferential S to D substitution in vertebrates are 0.549 and 0.597.

The differences in % of serine substitutions (S to D) of pS and npS are observed in poplar and rice (see Figure 4.6). Similarly to *S.bayanus* in yeast analysis, *A.lyrata* is very close to *A.thaliana*, thus does have a shorter time of divergence (~10 mya) and much lower percentage of serine substitution compared to poplar and rice. There seems to be a trend for preferential substitutions of pS in poplar and rice, but the signals are not always very strong (data not shown).

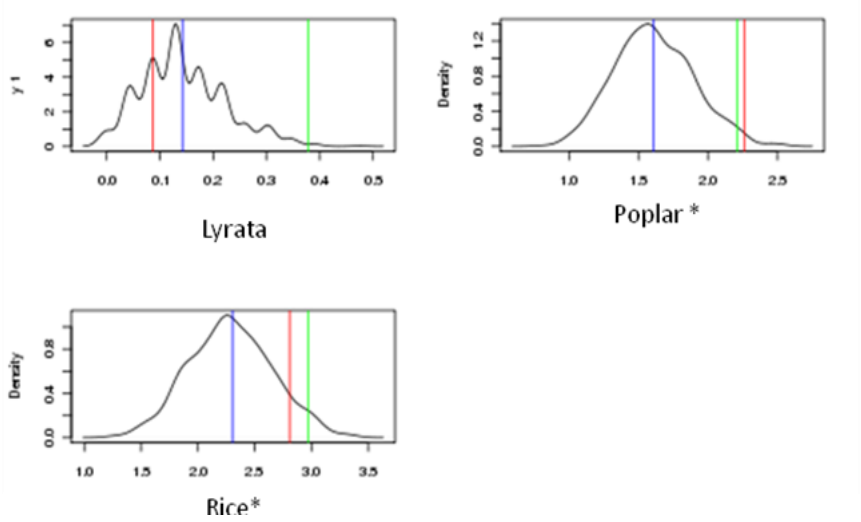


Figure 4.6: to Figure 4.2, this figure is for the run: **S to D substitution, blast_filter**, **full**, **re_1000 resampling**. Probabilities for such a preferential S to D substitution in plants are 0.157, 0.981 and 0.926.

4.3 Conclusion

In this study, using the yeast phosphoproteome, we have calculated the percentages of phospho-Serine substitutions (particularly to D, E and A) through pairwise alignments between *S. cerevisiae* genes and their orthologs in another six fungal species. The overall percentages (of different datasets and runs) for substitutions of pS to D, E and A are quite low (no higher than 5% for D and E, no higher than 8% for A), but still we have found enough support for a preferential substitution of pS when compared to non-phospho-Serines.

For most analyses in *S. bayanus*, the trend for a preferential substitution is not as strong as for other fungal species, which might be due to the short time of divergence (~20 mya) between *S. bayanus* and *S. cerevisiae* as well as the low overall percentages of substitutions (no higher than 0.5%). For the most distant (pre-WGD) species *Kluyveromyces lactis*, evidence for a preferential pS substitution is obvious, since this species has a relatively larger collection of S to D/E substitutions (or in general, enough time of divergence that allows more variations) compared to more

closely related species. For datasets where “aln_filter” is applied, evidence for a preferential substitution of pS is less strong, which might be due to the significant decrease in the number of substitutions in each dataset.

Nevertheless, our initial hypothesis for a preferential pS to D/E substitution has been supported by most of the analysis that we have conducted in this study using the 12HQ dataset from the budding yeast, which according to the analysis we have done in Chapter 2 can be used to accurately reveal certain properties of its phosphoproteome. In addition, a “disfavored” substitution of pS to A is uncovered. Therefore, pS is more frequently substituted to D/E and less frequently to A when compared to its non-modified amino acid counterpart.

The reason that we could not see any stronger evidence with human data might be that the current compendium of human phosphoproteome is far from complete. Therefore, the negative dataset that we collect might be “contaminated” with positive sites which have not been uncovered. If true, then it is no surprise that no evidence has been found with human data that could support our hypothesis. There seems to be some evidence for a preferential pS substitution in plants. Similarly to what we have found with *S. bayanus*, we could not draw clear conclusions from the substitution pattern in *A.lyrata* due to the short time of divergence with *A. thaliana*. But the substitutions in some cases in poplar and rice might be considered as evidence for a preferential substitution of pS in plants, even though the current collection of phosphorylation sites in *A. thaliana* is quite limited. When larger phosphoproteomic datasets are produced for human and A. *thaliana*, we should be able to understand better the substitution patterns in vertebrates and plants.

What is the reason for this preferential substitution (to D/E) of pS, at least in yeast? Why is this particular pattern selected/favored but other substitutions not? As we know, the amino-acid sequence motifs for phosphorylation are short (Gnad, Ren et al. 2007) and phosphorylation sites very frequently occur within rapidly evolving

unstructured regions (Iakoucheva, Radivojac et al. 2004). It is suggested by one of our previous studies with the yeast phosphoproteome that changes in phosphorylation sites might present a ready means of rapidly effecting the necessary re-wiring of gene regulatory networks. We have also shown in the same study that rapid evolution of these sites is linked to gene retention (Amoutzias, He et al. 2010). It also appears that phosphoproteins constitute the raw material for pathway rewiring and adaptation at various evolutionary rates (Chapter 2). So how can we link this preferential substitution that is observed in this study to the rewiring of gene regulatory networks, or even to species speciation/adaptation? It is also possible that the losses of “rewirable” interactions (null state interactions) between kinases and substituted pS (to D/E) are simply the result of adaptation. Substitutions of phosphorylated-Serines to alanine seem to be disfavored, when compared to the non-phosphorylated-Serines.

Recently, Kurmangaliyev *et al.* (Kurmangaliyev, Goland et al. 2011) has reported the same principal findings with similar methodologies, but on datasets of lower quality. Only two published phosphoproteomic datasets with no filters to address the issues of false positives and false negatives of p-sites are used in their study, while in our study, a high quality phosphorylation dataset is prepared from 12 publicly available yeast phosphoproteomic datasets. In addition, they have not addressed the issue of low-quality alignment that will raise some questions about their validity. In our analysis, the quality of alignment is well addressed by using homologous regions that are identified by BLAST.

In any case, our study should shed some light on the general properties of phosphorylated serines and help better understand the evolutionary pattern of phosphorylation sites.

4.4 Materials and Methods

4.4.1 Proteomes under investigation

The proteomes of seven fungal species were downloaded from YGOB (http://wolfe.gen.tcd.ie/ygob/, (Oheigeartaigh, Armisen et al. 2011). YGOB identifies the orthologous relationships among the proteins of these species. The seen species include the budding yeast, *Saccharomyces cerevisiae* (Goffeau, Barrell et al. 1996), four species that diverged after a whole genome duplication (WGD) (that occurred ~100 Mya, within the budding yeast lineage) and two fungi that diverged just prior to the WGD event. The four post-WGD species include *Vanderwaltozyma polysporus* (Scannell, Frank et al. 2007), *Naumovia castellii* (Cliften, Sudarsanam et al. 2003), *Candida glabrata* (Dujon, Sherman et al. 2004) and *Saccharomyces bayanus* (Kellis, Birren et al. 2004). The two fungi that diverged just prior to the WGD event include *Kluyveromyces lactis* (Dujon, Sherman et al. 2004) and *Zygosaccharomyces rouxii* (Souciet, Dujon et al. 2009). The evolutionary relationships of these seven species are depicted in the fungal phylogeny (shown in Figure 4.7).

To further investigate the substitution patterns in plants, the complete proteomes of the model plant species *Arabidopsis thaliana* (AGI 2000) together with its closest (fully sequenced and annotated) relative *Arabidopsis lyrata* (Hu, Pattyn et al. 2011) and two other plants, *Populus trichocarpa* (Tuskan, Difazio et al. 2006) and *Oryza sativa ssp. japonica* (Rice 2003) were downloaded from PLAZA (Proost, Van Bel et al. 2009). For investigating the vertebrate lineage, the proteomes of *Homo sapiens*, *Mus musculus* and *Gallus gallus* were retrieved from Ensembl (http://www.ensembl.org/; release 58, May 2010).

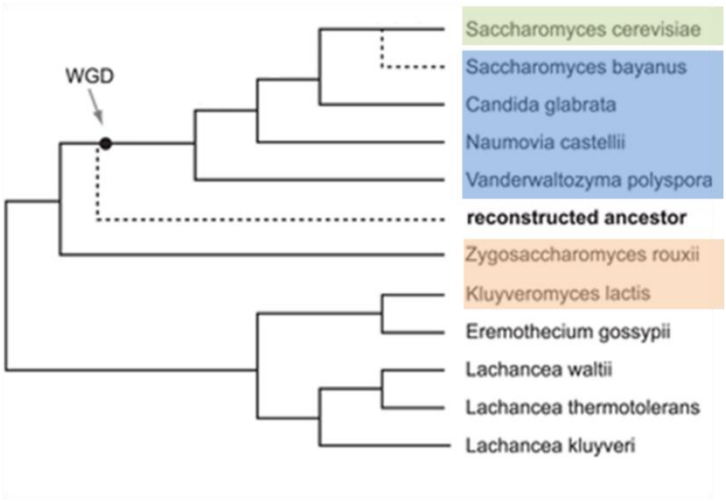


Figure 4.7: Phylogenetic relationships among all the 7 fungal species: budding yeast, *Saccharomyces Cerevisiae* (highlighted in green), two pre-WGD species, *Kluyveromyces lactis* and *Zygosaccharomyces rouxii* (highlighted in orange), and four post-WGD species, *Vanderwaltozyma polysporus*, *Naumovia castellii*, *Candida glabrata* and *Saccharomyces bayanus* (highlighted in blue) (taken from (Gordon, Byrne et al. 2009)).

4.4.2 Inference of orthologous relationships

Orthologous relationships in the fungal lineage were based on the homology pillars used in YGOB (Oheigeartaigh, Armisen et al. 2011) and on BLAST (bl2seq) whenever necessary (Altschul, Gish et al. 1990). Similarly, one-to-one orthologies for plants and vertebrates were retrieved from PLAZA (Proost, Van Bel et al. 2009) gene families (based on tribe-MCL and orthoMCL results) (Enright, Van Dongen et al. 2002; Li, Stoeckert et al. 2003) and Ensembl BioMart respectively (Flicek, Amode et al. 2011). Any cases of arbitrary relationships were further resolved by BLAST.

4.4.3 Experimentally determined p-sites

A dataset of high-quality phosphorylation sites (serines only) from the budding yeast (designated as 12HQ_pS) was prepared from 12 high-throughput experiments (Gruhler, Olsen et al. 2005; Chi, Huttenhower et al. 2007; Li, Gerber et al. 2007;

Albuquerque, Smolka et al. 2008; Bodenmiller, Campbell et al. 2008; Beltrao, Trinidad et al. 2009; Gnad, de Godoy et al. 2009; Holt, Tuch et al. 2009; Huber, Bodenmiller et al. 2009; Soufi, Kelstrup et al. 2009; Stark, Su et al. 2010) as shown in Chapter 2. This dataset is composed of 7913 p-sites from 2214 phosphoproteins. In addition, 117,575 non-phosphorylated serines, for which there is no evidence at all of phosphorylations were collected from 2370 12HQ phosphoproteins, thus comprising a negative dataset, (designated as 12HQ_npS1). The 12HQ_npS1 was further filtered with the Netphosyeast prediction algorithm, to generate a more stringent negative dataset, the 12HQ_npS2. With the help of Netphosyeast (that is considered the best performing algorithm for yeast motifs (Ingrell, Miller et al. 2007)) any strongly predicted (with prediction score > 0.7) phosphorylation sites were removed from the 12HQ_npS1 negative dataset, thus creating the 12HQ_npS2 dataset, composed of 100392 non-phosphorylated serines from 2368 12HQ phosphoproteins.

For the analysis on vertebrates, a collection of 8554 phospho-Serines from 2759 human phospho-proteins was extracted from Phospho.ELM (Dinkel, Chica et al. 2011), a database of experimentally verified phosphorylation sites in eukaryotic proteins. For the analysis on plants, 2519 phospho-Serines from 1511 phosphoproteins, identified by mass spectrometry in large-scale experiments were collected from the Arabidopsis Protein Phosphorylation Site Database PhosPhAt 3.0 (Durek, Schmidt et al. 2010).

4.4.4 Extraction of psite substitutions

Phosphorylation site substitutions between the orthologs of a reference species (*S. cerevisiae* for fungi, *H. sapiens* for vertebrates, *A. thaliana* for plants) and any other species of the same lineage were determined by pairwise alignment of the two orthologs, based on T-coffee (SI.Figure 4.1) (Notredame, Higgins et al. 2000). In particular, for the fungal lineage, one of the orthologs was always from S. *cerevisiae*

and the other ortholog was from one of the other six fungal species. Respectively, the same was applied for human phosphoproteins and their orthologs in mouse and chicken, and for the *Arabidopsis* phosphoproteins and their orthologs in A. lyrata, *P. trichocarpa* and *O. sativa*.

SI.Figure 4.1: Amino acid substitutions are extracted through pairwise alignments between orthologs. Here is an example of S to D substitution (pS is highlighted in black).

4.4.5 Considering sequence constraints

The rate of evolution is not homogeneous among the various proteins of an organisms and furthermore, the rate of evolution is not homogeneous within the various regions of a protein (Brown, Takayama et al. 2002). In particular, the evolution of protein domains is constrained by their 3D structure, whereas the intrinsically disordered regions are less constrained and thus evolve at higher rates. Interestingly, the majority of p-sites are found within such fast-evolving regions ((Iakoucheva, Radivojac et al. 2004) and Chapter 2). Thus, for determining the substitution patterns of phosphorylated and non-phosphorylated serines in our analyses, the above considerations need to be accounted for. In addition, pairwise alignment algorithms between two orthologs may attempt to align two regions that have lost any signal of homology. Thus, for the correct alignment of homologous sites, protein regions that have lost any signal of homology need to be filtered out. This additional filtering is performed with the help of blast, which identifies local alignments of homologous regions between any two orthologs.

For predicting intrinsically disordered regions, the VSL2B algorithm (which was ranked as the best performing relevant tool in CASP7) was used (Peng, Radivojac et

al. 2006). Thus, the phosphorylated serines dataset was further designated as “full” (e.g. 12HQ_pS_full) if it contained all phosphoserines, regardless of the structural constraints of the region they were located within, or designated as “disorder”, if the phosphorylated serines were located within intrinsically disordered regions. The same principle applied for the various negative datasets (npS1 and npS2) in fungi and also for all the other datasets in the vertebrate and plant lineages.

In a similar fashion, if a blast local-homology filter was applied (to identify homologous positions in a pairwise alignment between two orthologs) for a dataset, that dataset was further designated as “blast_filter”, whereas if no such local homology filter was applied, the dataset was designated as “no_filter”. Another filter that was applied to infer correct homology between two positions in a pairwise alignment (and thus ensure correct substitution patterns) was the use of neighborhood conservation. For a given amino acid, an 11 amino acid motif was created, with 5 residues to the left and 5 to the right of it. The particular amino acid (phosphorylated or non-phosphorylated serine) was further considered for analysis if at least 50% of the residues of that 11 amino acid motif were also conserved in the aligned region of the orthologous protein from another species. Datasets that underwent this type of filter were designated as “aln_filter”.

4.4.6 Statistical significance and probability for a preferential substitution

Two different types of resamplings were performed in order to reveal any differences in the substitution percentages (in particular from S to D/E) between phosphorylated and non-phosphorylated serines. In the first type of resamplings, designated as “original”, for each 12HQ phosphoprotein, serines from the negative dataset are randomly sampled with the same applied filters as those of the phosphorylation dataset (e.g. intrinsically disordered regions, netphosyeast filter, blast filter, >50%

conservation of 11 aa motif). They are designated as “ori” resampling and their number is equal to that of phosphorylated serines from the same phosphoprotein. Each resampling run is performed for all 12HQ phosphoproteins, with 1000 runs in total. For each resampling run, the overall percentage of S to D/E substitution is calculated and after 1000 runs, a distribution of % of non-phosphorylated S to D/E substitutions is generated. The figures with distributions are used to give an overview of the percentages of S to D/E substitution for pS, npS and reS. Protein structure is also taken into consideration. For example, if in protein A there exist N phosphorylated serines, within ID regions, then, N non-phosphorylated serines from ID regions of the same protein A are randomly sampled.

In the second type of resampling, designated as “re_1000”, for each resampling run, 1000 phosphorylated and 1000 non-phosphorylated serines are selected from the whole positive and the whole negative dataset (regardless of which proteins they are from). The same filters apply for both the positive and negative dataset (e.g. intrinsically disordered regions, netphosyeast filter, blast filter, 50% conservation of 11 aa motif).

The null hypothesis is that pS and npS have the same underlying percentage of S to D/E substitutions. Probability is calculated based on how many times (out of 1000) that pS had a higher % of S to D/E substitutions than the % of npS. A high probability indicates a preferential substitution in pS. Statistical significance is evaluated by Mann-Whitney U test in R.

4.5 Supporting Information

4.5.1 Summary for the full pS and npS collection

7913 p-sites (serines only, pS) from 2214 phosphoproteins are collected from a previous study (see Chapter 2). Two negative datasets (non-phosphorylated serines) npS1 and npS2 are generated for comparisons (see Materials and Methods, section 4.4.3). Statistics of phosphorylation sites in *S.cerevisiae* is shown in SI.Table 4.1. For the analysis on vertebrates, a collection of 8554 phospho-Serines from 2759 human phospho-proteins was extracted from Phospho.ELM (Dinkel, Chica et al. 2011). For the analysis on plants, 2519 phospho-Serines from 1511 phosphoproteins were collected from PhosPhAt 3.0 (Durek, Schmidt et al. 2010) (see SI.Table 4.1 and section 4.4.3 in Materials and Methods).

Species	Dataset	Num. of serines	Num. of proteins
S.cerevisiae	12HQ_pS	7913	2214
S.cerevisiae	12HQ_npS1	117574	2370
S.cerevisiae	12HQ_npS2	100392	2368
H.sapiens	phosphoELM	8554	2759
H.sapiens	H.negative	989141	21787
A.thaliana	PhosPhAt 3.0	2513	1507
A.thaliana	A.negative	1003687	27327

SI.Table 4.1: Summary of experimentally determined phosphorylation sites (serines only) and (the rest) non-phosphorylation sites in yeast, human and Arabidopsis, respectively.

A description of serine to all amino acid substitutions, for the pS and npS (both npS1 and npS2) datasets (for all variations of applied filters) are listed in SI.Table 4.2. When no filter ("no_filter") is applied to sequence conservation, the numbers of phosphorylated serines (pS) that are used to study substitutions are ranging from 7,400 to 7,700 for the "full" (full length) dataset, and from 7,100 to 7,400 for only the "disorder" dataset (serines from disordered regions), for different species respectively. The numbers of non-phosphorylated serines (npS) are ranging from 109,000 to 114,500 (npS1) and from 93,000 to 97,000 (npS2) for the "full" dataset; from 67,500 to 71,000 (npS1) and from 54,000 to 57,000 (npS2) for only the "disorder"s. When blast_filter (serines from BLAST homologous regions) is applied, the number of serines used to extract substitutions for different dataset is comparable to the "no_filter" dataset. However, when the more stringent strategy "aln_filter" (serines from well aligned regions) is applied, the numbers of substitutions for different datasets dropped significantly (SI.Table 4.2).

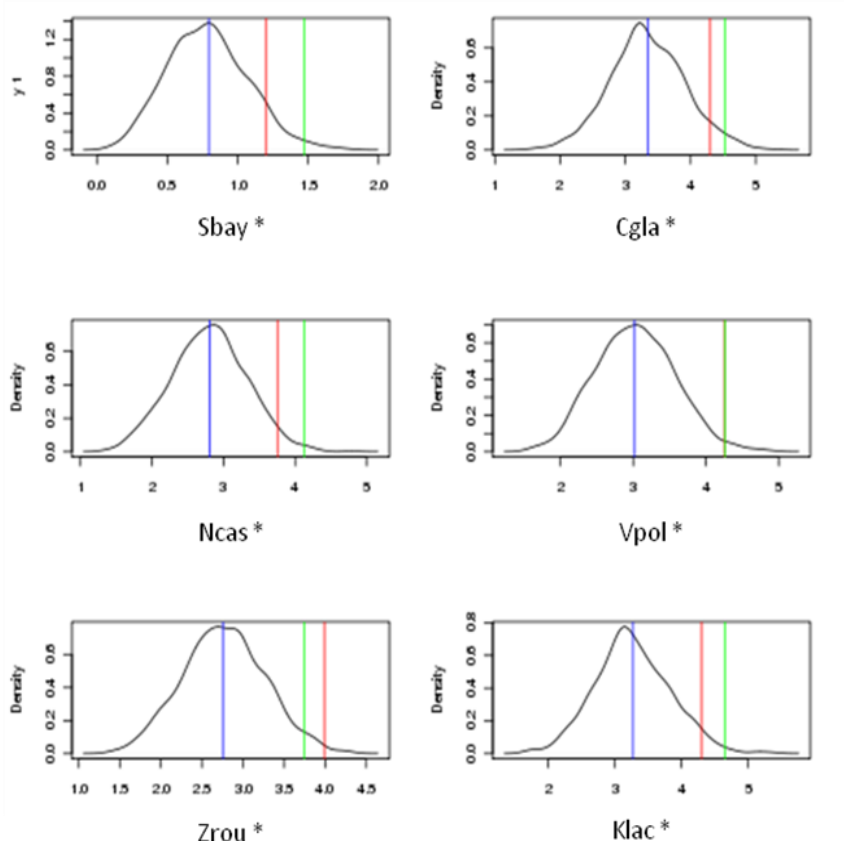
Filter	S	Data	Scer	Sbay	Cgla	Scas	Kpol	Zrou	Klac
no_filter	F	pS	7913	7451	7713	7754	7676	7684	7685
blast_filter				7395	7183	7186	7144	7234	7135
aln_filter				6487	2693	2922	2581	2735	2262
no_filter	D	pS	7334	6899	7161	7196	7129	7135	7132
blast_filter				6847	6633	6649	6605	6687	6586
aln_filter				5945	2196	2421	2105	2241	1781
no_filter	F	npS1	117574	109141	114203	114696	113852	114464	114456
blast_filter				108299	109479	108619	109149	110860	108749
aln_filter				98955	57950	60098	58188	61173	51862
no_filter	D	npS1	72725	67510	70983	71503	70831	71171	71201
blast_filter				66800	66571	66942	66801	67757	66013
aln_filter				57905	23551	25321	23628	25344	19840
no_filter	F	npS2	100392	93012	97590	97880	97191	97718	97694
blast_filter				92397	94182	93078	93755	95180	93432
aln_filter				85133	51910	53850	52270	54895	46529
no_filter	D	npS2	58310	54004	57010	57345	56815	57086	57105
blast_filter				53507	53893	53998	54016	54724	53337
aln_filter				46663	19670	21291	19889	21319	16567

SI.Table 4.2: The numbers of serine substitutions for pS and npS (both npS1 and npS2) in different runs, when protein structures, extra filters on sequence conservation are applied independently; **F** and **D** in the second column refers to full length and disorder, respectively.

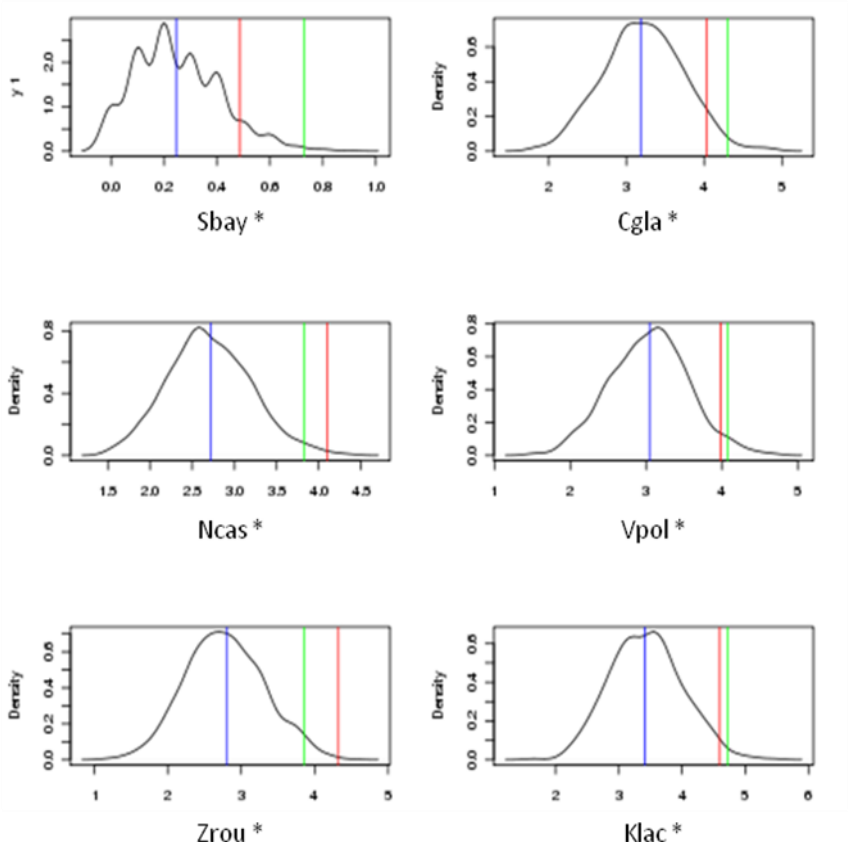
4.5.2 Substitutions for different datasets in fungi

4.5.2.1 Preferential substitution of pS to E

The overall percentages (%) of serine substitutions (S to E) for pS and npS (npS1 and npS2 separately) are listed in SI.Figure 4.2 and SI.Figure 4.3, for runs with “blast_filter” and for resampling type “re_1000”, respectively. There is significant difference between substitution ratio of pS and npS, for each single dataset (some results are not shown here). Again, the results for npS1 and npS2 are consistent, highlight the robustness of our analysis.



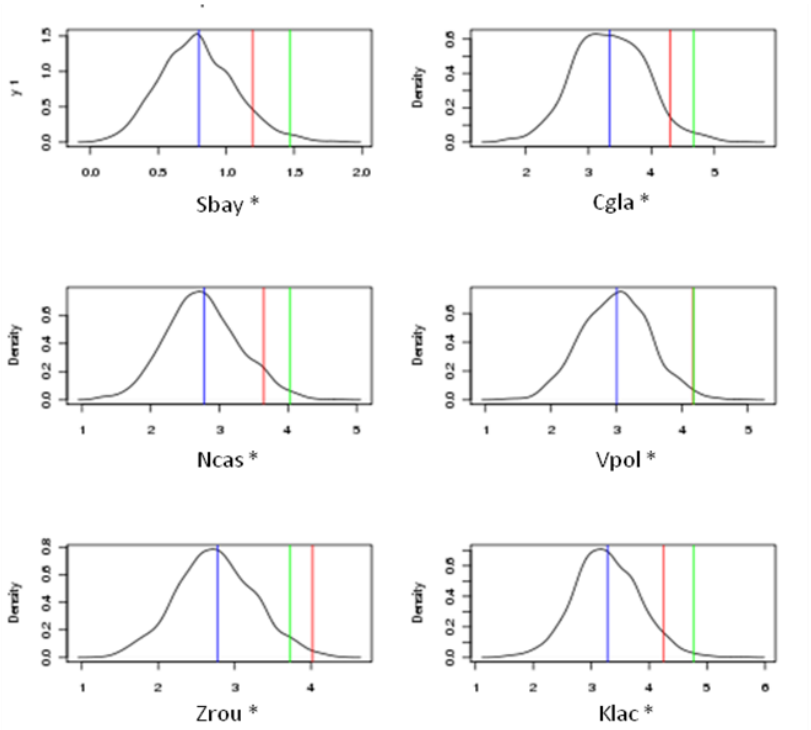
SI.Figure 4.2: Similarly to Figure 4.2, this figure is for the run: **S to E substitution**, **blast_filter**, **npS1**, **full**, **re_1000 resampling.** Probabilities for such a preferential S to E substitution of these 6 species are 0.883, 0.903, 0.99, 0.932, 0.999 and 0.971.



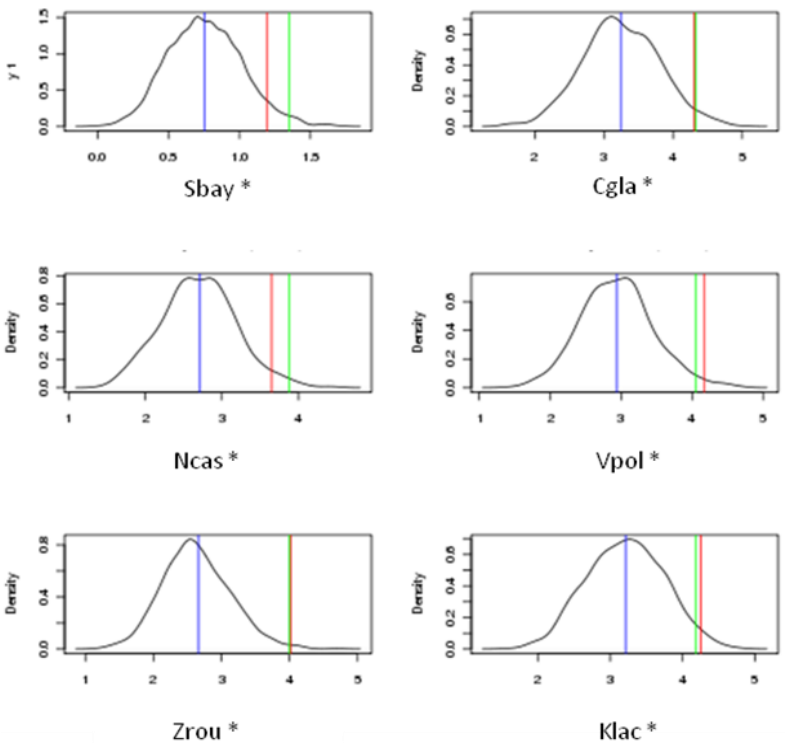
SI.Figure 4.3: Similarly to Figure 4.2, this figure is for the run: **S to E substitution**, **blast_filter**, **npS2**, **full**, **re_1000 resampling**. Probabilities for such a preferential S to E substitution of these 6 species are 0.895, 0.949, 0.996, 0.954, 0.999 and 0.973.

4.5.2.2 No control for the quality of alignment

The overall percentages (%) of serine substitutions (S to D) for pS and npS (npS1 and npS2) are listed in SI.Figure 4.4 and SI.Figure 4.5, for runs with without extra filters on protein sequence conservation (“no_filter”) and for resampling type “re_1000”, respectively.



SI.Figure 4.4: Similarly to Figure 4.2, this figure is for the run: **S to D substitution**, **no_filter**, **npS1**, **full**, **re_1000 resampling**. Probabilities for such a preferential S to D substitution of these 6 species are 0.899, 0.957, 0.942, 0.984, 0.992 and 0.952.

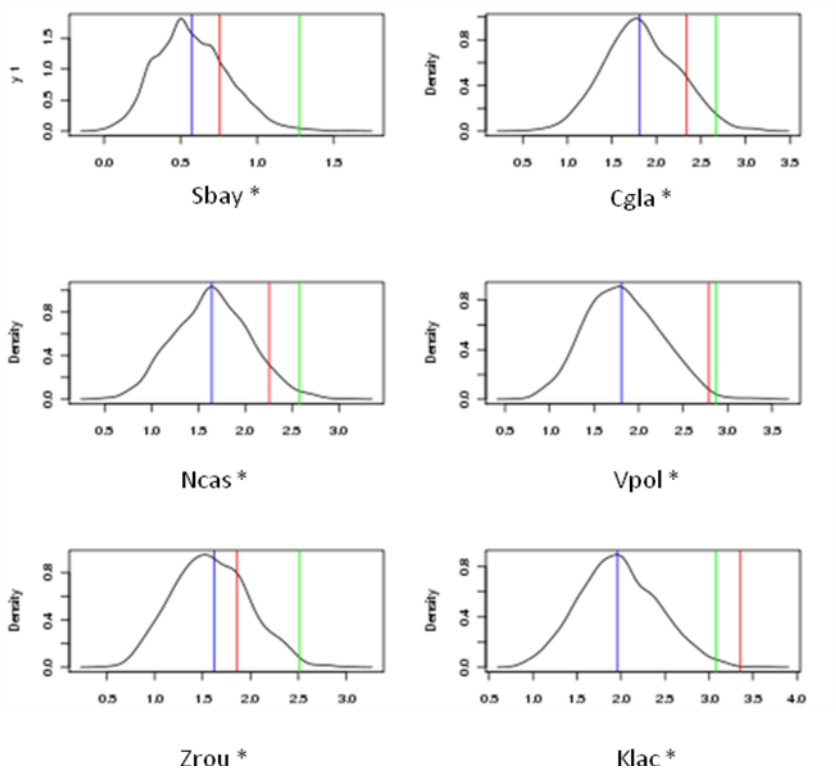


SI.Figure 4.5: Similarly to Figure 4.2, this figure is for the run: **S to D substitution**, **no_filter**, **npS2**, **full**, **re_1000 resampling**. Probabilities for such a preferential S to D substitution of these 6 species are 0.93, 0.961, 0.963, 0.983, 0.993 and 0.975.

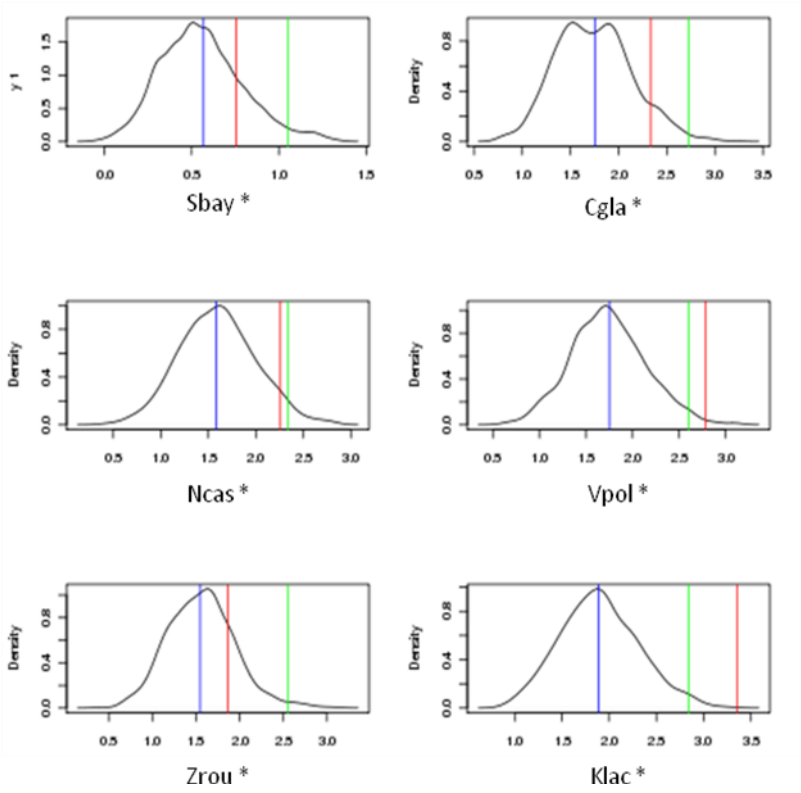
4.5.2.3 Control for over 50% conservation on sequence alignment

The overall percentages of serine substitutions for pS and npS are listed in SI.Figure 4.6 (npS1) and SI.Figure 4.7 (npS2) for the run with “aln_filter” (see Methods and Materials) using the “re_1000” resampling.

The results for npS1 and npS2 are consistent. There are significant differences in preference between pS and npS substitutions to D/E (one exception when only disorder regions are tested) even when very strict criteria is used. However, the trend for a preferential substitution is not as strong as in “blast_filter” runs. The most distantly related species *K.lactis* carries the strongest signals which might be due to its longest time of divergence.



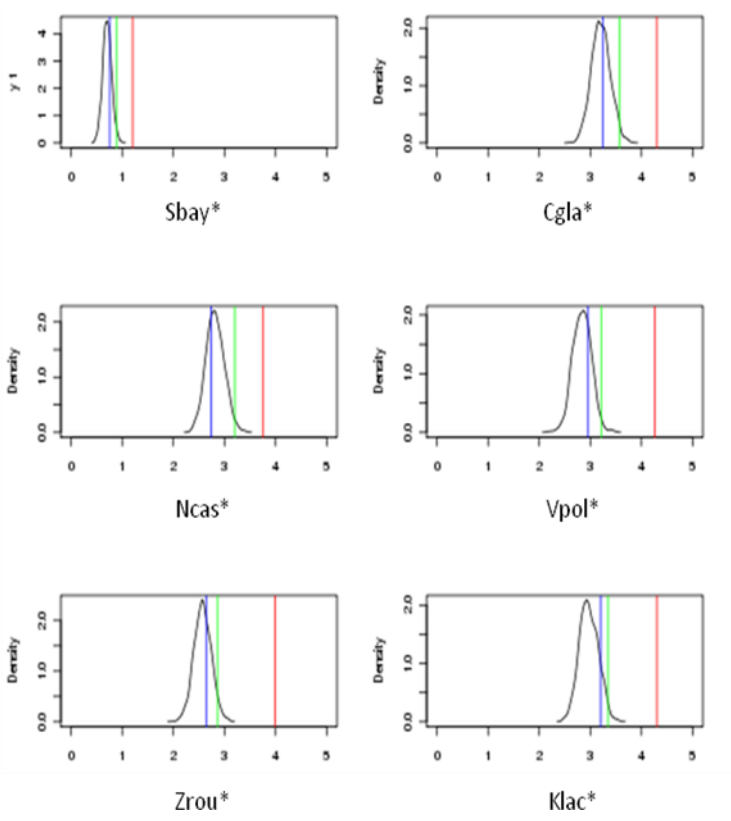
SI.Figure 4.6: Similarly to Figure 4.2, this figure is for the run: **S to D substitution**, **aln_filter**, **npS1**, **full**, **re_1000 resampling**. Probabilities for such a preferential S to D substitution of these 6 species are 0.78, 0.885, 0.929, 0.992, 0.74 and 0.998.



SI.Figure 4.7: Similarly to Figure 4.2, this figure is for the run: **S to D substitution**, **aln_filter**, **npS2**, **full**, **re_1000 resampling**. Probabilities for such a preferential S to D substitution of these 6 species are 0.808, 0.92, 0.942, 0.991, 0.786 and 0.999.

4.5.2.4 Original resampling

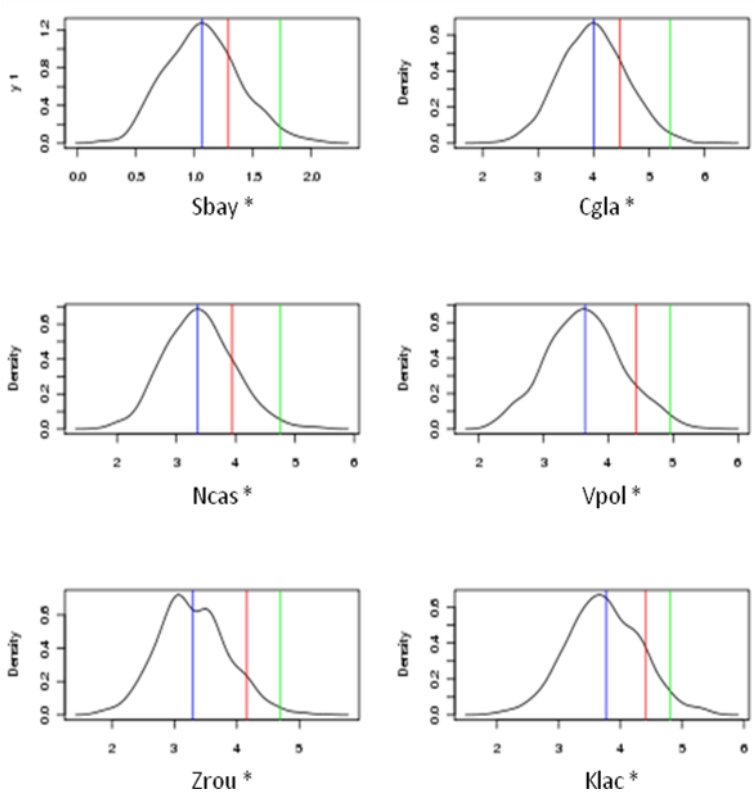
The overall percentages of serine substitutions for pS and npS (npS2 only) are listed in SI.Figure 4.8 for the run with “blast_filter” using the original resampling (see Methods and Materials). The differences between pS and npS is quite obvious, however, a deviation of average % of npS substitutions from reS dataset is also observed (blue line is not positioned at the middle of the distribution curve). The resampling (shown as the distribution curve) is slightly biased and might not represent the “real” negative dataset. Therefore, “re_1000” resamplings are used in this study to calculate the probability for a preferential substitution of pS.



SI.Figure 4.8: Similarly to Figure 4.2, this figure is for the run: **S to D substitution**, **blast_filter**, **npS2**, **full**, **ori_resampling**. Probabilities for such a preferential S to D substitution of these 6 species are are 1 for all.

4.5.2.5 Disordered regions only

The overall percentages of serine (disordered regions only, see Methods and Materials) substitutions for pS and npS2 are listed in SI.Figure 4.9 for the run with “blast_filter” using the “re_1000” resampling.



SI.Figure 4.9: Similarly to Figure 4.2, this figure is for the run: **S to D substitution**, **blast_filter**, **npS2**, **disorder**, **re_1000 resampling**. Probabilities for such a preferential S to D substitution of these 6 species are 0.723, 0.782, 0.833, 0.906, 0.926, and 0.891.

4.6 Author contributions

Y.H. conducted all of the bioinformatics analysis and wrote the manuscript under the supervision of G.A. and Y.V.D.P..

5 Concluding remarks

5.1 General conclusions.

General properties of the yeast phosphoproteome have been investigated in Chapter 2. A high-quality curated phosphoproteomic compendium is assembled from 12 publicly available phosphoproteomic datasets. We have provided a HQ compendium filtered of noise, by applying very stringent criteria that filter out both technical false-positives and low-stoichiometry off-target phosphorylations. This HQ dataset of p-sites has been used to study the general properties of the yeast phosphoproteome in a comprehensive way. Our analyses show that the compendium may well be approaching saturation in terms of identifying all yeast phosphoproteins, but it is far from complete in terms of identifying all the p-sites on those proteins. One interesting finding that kinases may mistakenly phosphorylate serines that are in the neighborhood of a cognate phosphorylation motif may particularly help understand better how kinases function.

The impact of post-translational regulation, in particular phosphorylation, on eukaryotic genome evolution has been studied in Chapter 3. We reported for the first time in literature that phosphorylation affects the retention of duplicated genes (especially after genome duplication) in the fungal lineages. It is shown in our study that proteins retained in duplicate are subject to more post-translational control and particularly to more phosphorylation than returned to single status proteins. Post-translational regulation repeatedly affected the future of gene duplicates in the various post-WGD yeast lineages, suggesting that gene retention is, to some extent, pre-determined. This finding may help better understand the basic mechanisms of gene retention and genome evolution.

In addition, the amino acid substitution pattern of phosphorylation sites has been investigated in Chapter 4. This study observed that phosphorylated serines, compared to their non-modified counterparts, tend to substitute more frequently than expected, to amino acids (aspartic acid, glutamic acid) that shared similar phosphomimetic properties and less frequently than expected to alanine, which resembles non-phosphorylation status. This study shed light on the general properties of phosphorylated serines and helped to better understand the evolutionary pattern of phosphorylation sites. In addition, this information may be exploited in the near future by algorithms that attempt to predict phosphorylation sites.

5.2 The dark side of the moon: can we rely on the data?

As stated in Chapter 2, concerns have been raised about the detection of "noisy" non-functional p-sites from different high throughput experiments, which is due to the high sensitivity of MS instruments. At the same time, experts in the field raise concerns about the incompleteness of the current phosphoproteome compendium. Investigating the properties of yeast phosphoproteome and studying the impact of phosphorylation on genome evolution with this "noisy" dataset (see Chapter 2) would be similar to a Chinese proverb: visualize the whole animal by looking at one spot on a leopard. Therefore, some very strict filters have been applied on the original datasets extracted from HTP experiments and a high-confidence subset has been obtained. We have investigated the properties of the yeast phosphoproteome in a thorough manner, by repeating analyses on a much larger subset as well as a literature-curated low throughput dataset.

Recent and impressive advances in genomic sequencing technology and development of NGS based tools and software make us wonder whether the time for an equivalent "revolution" in proteomics is "round the corner". When researchers are trying to investigate the genetic causes of a certain disease, it is not (always)

sufficient to know which genes and/or which parts of the corresponding DNA sequences are present and involved. More importantly, one needs to have a clear picture of which proteins are responsible for biological processes of interest, such as growth and metabolism, and disease. We are counting on the next revolution in proteomics technologies. Work needs to be done in establishing techniques into a new basis technology and developing new application methods based on this new technology.

5.3 The importance of being critical and taking advantage of whatever is there.

There is a wide and increasingly intense interest in the field of phosphoproteomics. Only recently, with the advent of high-throughput phosphoproteomics data, we can address a few questions that could not be properly answered before. Despite the weaknesses of current phosphoproteomics technology (discussed in Chapter 2 and paragraphs above), we have addressed in a very rigorous manner many interesting and important questions. However, one always needs to be aware of and critical of the available published data. This is also why we think that the study of evaluation and properties of the yeast phosphoproteome is interesting and important to a wide audience (both experimental and computational biologists). One should always think how reliable a dataset is (check the properties of the dataset) and how to “clean” the data (using bioinformatics tools) before using it. The issues of false positives and false negatives of high throughput phosphorylation data have been discussed and addressed at our best in Chapter 2.

For experimental biologists, a better understanding of the properties of yeast phosphoproteome would help improve the protocols, which might lead to advances of p-site detection (which will give computational scientists more material to work on). When more (new) phosphoproteomics data are available, we would obtain a deeper

understanding of the general properties of this post-translational modification as well as its impact on genome evolution. For computational biologists, general features of phosphoproteome and phosphorylation sites should facilitate the development of tools/models to process and interpretation of the data. For example, we know from the literature and our own study (Chapter 2) that most phosphorylation sites (~91%) are found in disordered regions (Landry, Levy et al. 2009), while only ~54% of non-phosphorylated sites are from disordered regions. Also the 7-amino-acids motif (3 amino acids on the left and 3 amino acids on the right) of each pSite shows certain level of conservation (Gnad, Ren et al. 2007), and a preferential substitution of phosphorylated-Serines compared to np-Serines is observed (Chapter 4) within our curated datasets (also confirmed by Kurmangaliyev *et al.* (Kurmangaliyev, Goland et al. 2011)). The signal for this pattern is there, not always strong. But still, all these information could be taken into consideration when a computer science engineer is trying to develop/optimize his tools/models for p-site prediction.

5.4 An excellent model organism.

The model organism *Saccharomyces cerevisiae* was chosen due to its simplicity, many publicly available functional genomics data and a wealth of information about the fungal phylogeny. When researchers (experimental biologists) look for an organism to use in their studies, they look for several traits: size of the genome, generation time, accessibility, manipulation, genetics, conservation of mechanisms, and potential economic benefit. Indeed, *S. cerevisiae* is an excellent model because it has a very small (and compact) genome, one whole genome duplication, many fully sequenced relatives (closed and distant) and a huge collection and high coverage of interactions (both experimental and predicted) involved in gene regulation (see Table 6.1 and Figure 6.1, e.g. protein-protein interactions).

Features	Yeast	Fruit fly	Human	Arabidopsis
Genome size (Mb)	12.1	180	3200	125
Approximate num. of genes	5800	13600	35000	25500
Gene density (avg. num. per Mb)	479	76	11	204
Avg. introns per gene	0.04	3	9	4
Num. of whole genome duplication	1	0	2	3
Num. of genes in PPI network	4874	7254	8655	1591
Num. of interactions in PPI network	36391	24799	26783	2588

Table6.1: Statistics of yeast, fruit fly, human and Arabidopsis genomes, protein-protein interaction data is from BioGRID v2.0.60.

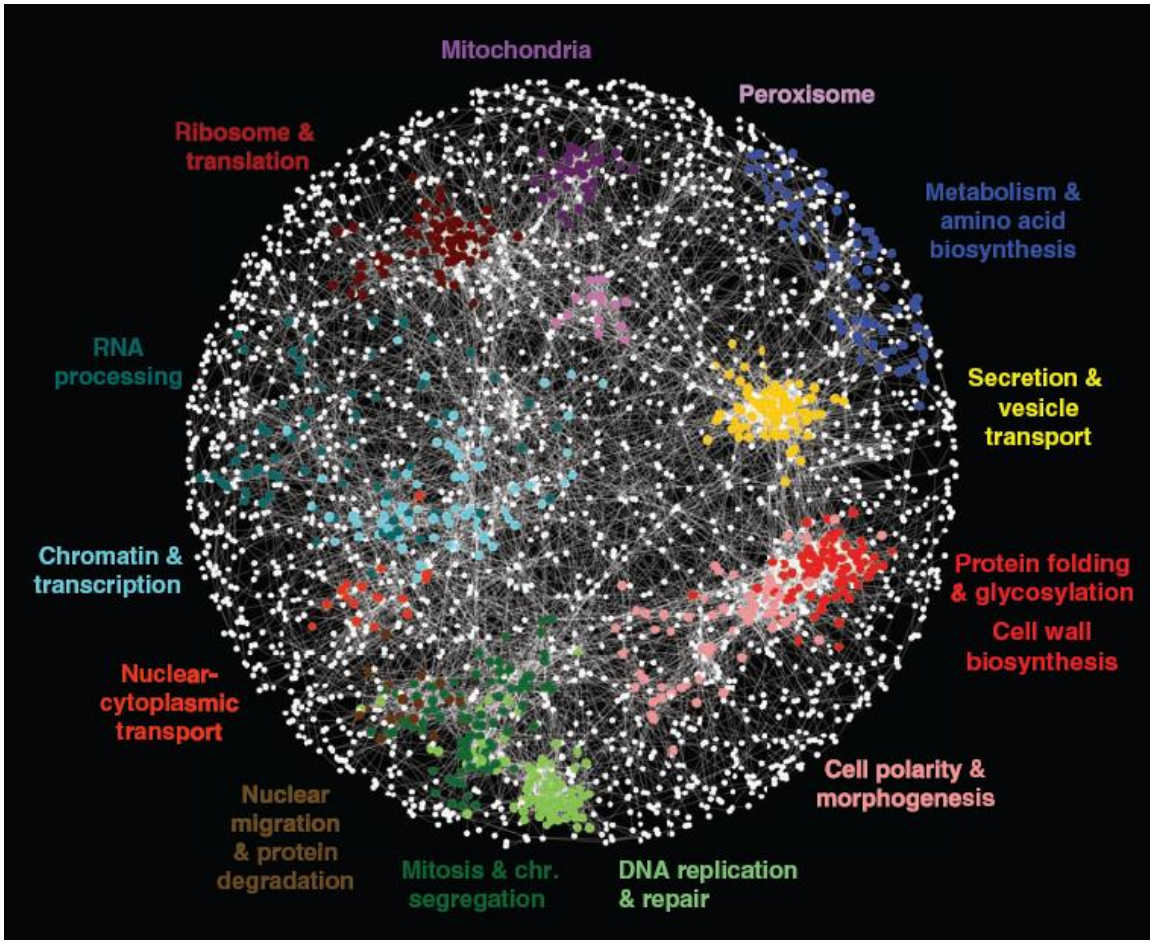


Figure 6.1: The synthetic gene array network as described by Costanzo *et al* (Costanzo, Baryshnikova et al. 2010). Each point represents a functional gene and each edge a functional connection, and the colored regions indicate subnetworks with similar GO process annotations.

Especially for systems biology (the study of interactions among genes, proteins, regulatory elements, metabolites using models and/or networks), the yeast *Saccharomyces cerevisiae* is still one of the best eukaryotic model organism where a concentrated research effort (in terms of genomic and functional data) has been put

into and significant advance in integrating biological data and interpreting gene networks has been made (in terms of development of computational tools and models) (see Figure 6.1). Although the yeast genome is hundreds of times smaller than that of the human, it displays considerable complexity with regard to diversity of gene expression.

The networks in systems biology might never completely represent the actual biological system (Yuan, Galbraith et al. 2008). With the advance in the study (genomics, functional genomics, and gene regulatory network) with yeast, normalization, standardization and visualization of all these biological data, modeling of gene regulatory networks will be learnt. Discoveries made in yeast will provide insight into the workings of other organisms (human and plants).

With the high coverage (phospho-)proteome in yeast, we can address interesting questions about its general properties and impact on genome evolution (as shown in Chapter 2, 3 and 4) that cannot be addressed properly at this moment in human and plants (Chapter 4). With the current MS-based technologies, characterizing the complete proteome (refers to the study of the full set of proteins in a cell type or tissue and the changes during various conditions) of a multicellular organism is still very challenging compared to the study on a unicellular yeast species. Many more experiments need to be conducted for different tissues and under various conditions to reach a decent coverage of its proteome. Therefore, for the study of post-translational regulation and its impact on genome evolution, *S.cerevisiae* is currently the one with the most data. When advances have been made in proteomics and much higher coverage phosphoproteome of human/Arabidopsis has been obtained, the questions that are asked at the beginning of this thesis, such as the relation between post-translational regulation and genome evolution of eukaryotes, could be answered in a much better and more comprehensive way in other eukaryotic lineages.

6 References

Aebersold, R. and M. Mann (2003). "Mass spectrometry-based proteomics." Nature **422**(6928): 198-207.

AGI (2000). "Analysis of the genome sequence of the flowering plant Arabidopsis thaliana." Nature **408**(6814): 796-815.

Albuquerque, C. P., M. B. Smolka, et al. (2008). "A multidimensional chromatography technology for in-depth phosphoproteome analysis." Mol Cell Proteomics **7**(7): 1389-96.

Amanchy, R., D. E. Kalume, et al. (2005). "Phosphoproteome analysis of HeLa cells using stable isotope labeling with amino acids in cell culture (SILAC)." J Proteome Res **4**(5): 1661-71.

Amores, A., A. Force, et al. (1998). "Zebrafish hox clusters and vertebrate genome evolution." Science **282**(5394): 1711-4.

Amoutzias, G. D., Y. He, et al. (2010). "Posttranslational regulation impacts the fate of duplicated genes." Proc Natl Acad Sci U S A **107**(7): 2967-71.

Amoutzias, G. D., D. L. Robertson, et al. (2004). "Convergent evolution of gene networks by single-gene duplications in higher eukaryotes." EMBO Rep **5**(3): 274-9.

Amoutzias, G. D., A. S. Veron, et al. (2007). "One billion years of bZIP transcription factor evolution: conservation and change in dimerization and DNA-binding site specificity." Mol Biol Evol **24**(3): 827-35.

Anthis, N. (2009). On Mimicking Phosphotyrosine.

Ashburner, M., C. A. Ball, et al. (2000). "Gene ontology: tool for the unification of biology. The Gene Ontology Consortium." Nat Genet **25**(1): 25-9.

Balaji, S., L. M. Iyer, et al. (2008). "Comparison of transcription regulatory interactions inferred from high-throughput methods: what do they reveal?" Trends Genet **24**(7): 319-23.

Barford, D., A. K. Das, et al. (1998). "The structure and mechanism of protein phosphatases: insights into catalysis and regulation." Annu Rev Biophys Biomol Struct **27**: 133-64.

Batada, N. N., T. Reguly, et al. (2006). "Stratus not altocumulus: a new view of the yeast protein interaction network." PLoS Biol **4**(10): e317.

Beausoleil, S. A., M. Jedrychowski, et al. (2004). "Large-scale characterization of HeLa cell nuclear phosphoproteins." Proc Natl Acad Sci U S A **101**(33): 12130-5.

Belle, A., A. Tanay, et al. (2006). "Quantification of protein half-lives in the budding yeast proteome." Proc Natl Acad Sci U S A **103**(35): 13004-9.

Beltrao, P., J. C. Trinidad, et al. (2009). "Evolution of phosphoregulation: comparison of phosphorylation patterns across yeast species." PLoS Biol **7**(6): e1000134.

Berg, J. M., J. L. Tymoczko, et al. (2002). Biochemistry. New York, W. H. Freeman.

Birchler, J. A. and R. A. Veitia (2007). "The gene balance hypothesis: from classical genetics to modern genomics." Plant Cell **19**(2): 395-402.

Blanc, G. and K. H. Wolfe (2004). "Widespread paleopolyploidy in model plant species inferred from age distributions of duplicate genes." Plant Cell **16**(7): 1667-78.
Bodenmiller, B., D. Campbell, et al. (2008). "PhosphoPep--a database of protein phosphorylation sites in model organisms." Nat Biotechnol **26**(12): 1339-40.
Bodenmiller, B., L. N. Mueller, et al. (2007). "Reproducible isolation of distinct, overlapping segments of the phosphoproteome." Nat Methods **4**(3): 231-7.
Brown, C. J., S. Takayama, et al. (2002). "Evolutionary rate heterogeneity in proteins with long disordered regions." J Mol Evol **55**(1): 104-10.
Brown, T. A. (2002). Genomes. Oxford, Wiley-Liss.
Byrne, K. P. and K. H. Wolfe (2005). "The Yeast Gene Order Browser: combining curated homology and syntenic context reveals gene fate in polyploid species." Genome Res **15**(10): 1456-61.
Cagney, G., S. Amiri, et al. (2003). "In silico proteome analysis to facilitate proteomics experiments using mass spectrometry." Proteome Sci **1**(1): 5.
Casneuf, T., S. De Bodt, et al. (2006). "Nonrandom divergence of gene expression following gene and genome duplications in the flowering plant Arabidopsis thaliana." Genome Biol **7**(2): R13.
Chang, C. and R. C. Stewart (1998). "The two-component system. Regulation of diverse signaling pathways in prokaryotes and eukaryotes." Plant Physiol **117**(3): 723-31.
Chi, A., C. Huttenhower, et al. (2007). "Analysis of phosphorylation sites on proteins from Saccharomyces cerevisiae by electron transfer dissociation (ETD) mass spectrometry." Proc Natl Acad Sci U S A **104**(7): 2193-8.
Cliften, P., P. Sudarsanam, et al. (2003). "Finding functional features in Saccharomyces genomes by phylogenetic footprinting." Science **301**(5629): 71-6.
Comai, L. (2005). "The advantages and disadvantages of being polyploid." Nat Rev Genet **6**(11): 836-46.
Costanzo, M., A. Baryshnikova, et al. (2010). "The genetic landscape of a cell." Science **327**(5964): 425-31.
Cox, J. and M. Mann (2010). "Quantitative, High-Resolution Proteomics for Data-Driven Systems Biology." Annu Rev Biochem.
Davis, J. C. and D. A. Petrov (2005). "Do disparate mechanisms of duplication add similar genes to the genome?" Trends Genet **21**(10): 548-51.
de Godoy, L. M., J. V. Olsen, et al. (2008). "Comprehensive mass-spectrometry-based proteome quantification of haploid versus diploid yeast." Nature **455**(7217): 1251-4.
Dehal, P. and J. L. Boore (2005). "Two rounds of whole genome duplication in the ancestral vertebrate." PLoS Biol **3**(10): e314.
Dietrich, F. S., S. Voegeli, et al. (2004). "The Ashbya gossypii genome as a tool for mapping the ancient Saccharomyces cerevisiae genome." Science **304**(5668): 304-7.
Dinkel, H., C. Chica, et al. (2011). "Phospho.ELM: a database of phosphorylation sites--update 2011." Nucleic Acids Res **39**(Database issue): D261-7.
Duggan, D. J., M. Bittner, et al. (1999). "Expression profiling using cDNA microarrays." Nat Genet **21**(1 Suppl): 10-4.

Dujon, B., D. Sherman, et al. (2004). "Genome evolution in yeasts." Nature **430**(6995): 35-44.
Dunker, A. K., C. J. Brown, et al. (2002). "Intrinsic disorder and protein function." Biochemistry **41**(21): 6573-82.
Dunker, A. K. and Z. Obradovic (2001). "The protein trinity--linking function and disorder." Nat Biotechnol **19**(9): 805-6.
Durek, P., R. Schmidt, et al. (2010). "PhosPhAt: the Arabidopsis thaliana phosphorylation site database. An update." Nucleic Acids Res **38**(Database issue): D828-34.
Enright, A. J., S. Van Dongen, et al. (2002). "An efficient algorithm for large-scale detection of protein families." Nucleic Acids Res **30**(7): 1575-84.
Evans, B. J., D. B. Kelley, et al. (2005). "Evolution of RAG-1 in polyploid clawed frogs." Mol Biol Evol **22**(5): 1193-207.
Fawcett, J. A., S. Maere, et al. (2009). "Plants with double genomes might have had a better chance to survive the Cretaceous–Tertiary extinction event." Proc Natl Acad Sci U S A.
Flicek, P., M. R. Amode, et al. (2011). "Ensembl 2011." Nucleic Acids Res **39**(Database issue): D800-6.
Freeling, M. and B. C. Thomas (2006). "Gene-balanced duplications, like tetraploidy, provide predictable drive to increase morphological complexity." Genome Res **16**(7): 805-14.
Furihata, T., K. Maruyama, et al. (2006). "Abscisic acid-dependent multisite phosphorylation regulates the activity of a transcription activator AREB1." Proc Natl Acad Sci U S A **103**(6): 1988-93.
Garske, A. L., U. Peters, et al. (2011). "Chemical genetic strategy for targeting protein kinases based on covalent complementarity." Proc Natl Acad Sci U S A.
Ghaemmaghami, S., W. K. Huh, et al. (2003). "Global analysis of protein expression in yeast." Nature **425**(6959): 737-41.
Giaever, G., A. M. Chu, et al. (2002). "Functional profiling of the Saccharomyces cerevisiae genome." Nature **418**(6896): 387-91.
Gnad, F., L. M. de Godoy, et al. (2009). "High-accuracy identification and bioinformatic analysis of in vivo protein phosphorylation sites in yeast." Proteomics **9**(20): 4642-52.
Gnad, F., S. Ren, et al. (2007). "PHOSIDA (phosphorylation site database): management, structural and evolutionary investigation, and prediction of phosphosites." Genome Biol **8**(11): R250.
Goffeau, A., B. G. Barrell, et al. (1996). "Life with 6000 genes." Science **274**(5287): 546, 563-7.
Gordon, J. L., K. P. Byrne, et al. (2009). "Additions, losses, and rearrangements on the evolutionary route from a reconstructed ancestor to the modern Saccharomyces cerevisiae genome." PLoS Genet **5**(5): e1000485.
Gregory, T. R. (2001). "Coincidence, coevolution, or causation? DNA content, cell size, and the C-value enigma." Biol Rev Camb Philos Soc **76**(1): 65-101.
Gruhler, A., J. V. Olsen, et al. (2005). "Quantitative phosphoproteomics applied to the yeast pheromone signaling pathway." Mol Cell Proteomics **4**(3): 310-27.
Gsponer, J., M. E. Futschik, et al. (2008). "Tight regulation of unstructured proteins: from transcript synthesis to protein degradation." Science **322**(5906): 1365-8.

Guan, Y., M. J. Dunham, et al. (2007). "Functional analysis of gene duplications in Saccharomyces cerevisiae." Genetics **175**(2): 933-43.
Gunawardena, J. (2005). "Multisite protein phosphorylation makes a good threshold but can be a poor switch." Proc Natl Acad Sci U S A **102**(41): 14617-22.
Hakes, L., J. W. Pinney, et al. (2007). "All duplicates are not equal: the difference between small-scale and genome duplication." Genome Biol **8**(10): R209.
Harbison, C. T., D. B. Gordon, et al. (2004). "Transcriptional regulatory code of a eukaryotic genome." Nature **431**(7004): 99-104.
He, X. and J. Zhang (2005). "Gene complexity and gene duplicability." Curr Biol **15**(11): 1016-21.
He, X. and J. Zhang (2005). "Rapid subfunctionalization accompanied by prolonged and substantial neofunctionalization in duplicate gene evolution." Genetics **169**(2): 1157-64.
Hedges, S. B., J. E. Blair, et al. (2004). "A molecular timescale of eukaryote evolution and the rise of complex multicellular life." BMC Evol Biol **4**: 2.
Holt, L. J., B. B. Tuch, et al. (2009). "Global analysis of Cdk1 substrate phosphorylation sites provides insights into evolution." Science **325**(5948): 1682-6.
Hsia, C. C. and W. McGinnis (2003). "Evolution of transcription factor function." Curr Opin Genet Dev **13**(2): 199-206.
Hu, T. T., P. Pattyn, et al. (2011). "The Arabidopsis lyrata genome sequence and the basis of rapid genome size change." Nat Genet **43**(5): 476-81.
Huber, A., B. Bodenmiller, et al. (2009). "Characterization of the rapamycin-sensitive phosphoproteome reveals that Sch9 is a central coordinator of protein synthesis." Genes Dev **23**(16): 1929-43.
Hunter, T. (2000). "Signaling--2000 and beyond." Cell **100**(1): 113-27.
Iakoucheva, L. M., C. J. Brown, et al. (2002). "Intrinsic disorder in cell-signaling and cancer-associated proteins." J Mol Biol **323**(3): 573-84.
Iakoucheva, L. M., P. Radivojac, et al. (2004). "The importance of intrinsic disorder for protein phosphorylation." Nucleic Acids Res **32**(3): 1037-49.
Ingrell, C. R., M. L. Miller, et al. (2007). "NetPhosYeast: prediction of protein phosphorylation sites in yeast." Bioinformatics **23**(7): 895-7.
Jaillon, O., J. M. Aury, et al. (2004). "Genome duplication in the teleost fish Tetraodon nigroviridis reveals the early vertebrate proto-karyotype." Nature **431**(7011): 946-57.
Johnson, S. A. and T. Hunter (2004). "Phosphoproteomics finds its timing." Nat Biotechnol **22**(9): 1093-4.
Kellis, M., B. W. Birren, et al. (2004). "Proof and evolutionary analysis of ancient genome duplication in the yeast Saccharomyces cerevisiae." Nature **428**(6983): 617-24.
Kersten, B., G. K. Agrawal, et al. (2009). "Plant phosphoproteomics: an update." Proteomics **9**(4): 964-88.
Kunin, V., J. B. Pereira-Leal, et al. (2004). "Functional evolution of the yeast protein interaction network." Mol Biol Evol **21**(7): 1171-6.
Kurmangaliyev, Y. Z., A. Goland, et al. (2011). "Evolutionary patterns of phosphorylated serines." Biol Direct **6**: 8.
Landry, C. R., E. D. Levy, et al. (2009). "Weak functional constraints on phosphoproteomes." Trends Genet **25**(5): 193-7.

Lang, D., B. Weiche, et al. (2010). "Genome-wide phylogenetic comparative analysis of plant transcriptional regulation: a timeline of loss, gain, expansion, and correlation with complexity." Genome Biol Evol **2**: 488-503.
Latchman, D. S. (2001). "Transcription factors: bound to activate or repress." Trends Biochem Sci **26**(4): 211-3.
Lee, T. I., N. J. Rinaldi, et al. (2002). "Transcriptional regulatory networks in Saccharomyces cerevisiae." Science **298**(5594): 799-804.
Li, J., X. Q. Wang, et al. (2000). "Regulation of abscisic acid-induced stomatal closure and anion channels by guard cell AAPK kinase." Science **287**(5451): 300-3.
Li, L., C. J. Stoeckert, Jr., et al. (2003). "OrthoMCL: identification of ortholog groups for eukaryotic genomes." Genome Res **13**(9): 2178-89.
Li, W. H., J. Yang, et al. (2005). "Expression divergence between duplicate genes." Trends Genet **21**(11): 602-7.
Li, X., S. A. Gerber, et al. (2007). "Large-scale phosphorylation analysis of alpha-factor-arrested Saccharomyces cerevisiae." J Proteome Res **6**(3): 1190-7.
Lienhard, G. E. (2008). "Non-functional phosphorylations?" Trends Biochem Sci **33**(8): 351-2.
Lin, M. H., T. L. Hsu, et al. (2009). "Phosphoproteomics of Klebsiella pneumoniae NTUH-K2044 reveals a tight link between tyrosine phosphorylation and virulence." Mol Cell Proteomics **8**(12): 2613-23.
Linding, R., L. J. Jensen, et al. (2007). "Systematic discovery of in vivo phosphorylation networks." Cell **129**(7): 1415-26.
Lodish, H., A. Berk, et al. (2000). Molecular Cell Biology. New York, W. H. Freeman.
Lynch, M. and J. S. Conery (2000). "The evolutionary fate and consequences of duplicate genes." Science **290**(5494): 1151-5.
Macek, B., F. Gnad, et al. (2008). "Phosphoproteome analysis of E. coli reveals evolutionary conservation of bacterial Ser/Thr/Tyr phosphorylation." Mol Cell Proteomics **7**(2): 299-307.
Macek, B., I. Mijakovic, et al. (2007). "The serine/threonine/tyrosine phosphoproteome of the model bacterium Bacillus subtilis." Mol Cell Proteomics **6**(4): 697-707.
Maere, S., S. De Bodt, et al. (2005). "Modeling gene and genome duplications in eukaryotes." Proc Natl Acad Sci U S A **102**(15): 5454-9.
Maere, S., K. Heymans, et al. (2005). "BiNGO: a Cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks." Bioinformatics **21**(16): 3448-9.
Manning, G., G. D. Plowman, et al. (2002). "Evolution of protein kinase signaling from yeast to man." Trends Biochem Sci **27**(10): 514-20.
McLysaght, A., K. Hokamp, et al. (2002). "Extensive genomic duplication during early chordate evolution." Nat Genet **31**(2): 200-4.
Mintseris, J. and Z. Weng (2005). "Structure, function, and evolution of transient and obligate protein-protein interactions." Proc Natl Acad Sci U S A **102**(31): 10930-5.
Misteli, T. (2007). "Beyond the sequence: cellular organization of genome function." Cell **128**(4): 787-800.
Miyata, T. and H. Suga (2001). "Divergence pattern of animal gene families and relationship with the Cambrian explosion." Bioessays **23**(11): 1018-27.

Mok, J., P. M. Kim, et al. (2010). "Deciphering protein kinase specificity through large-scale analysis of yeast phosphorylation site motifs." Sci Signal **3**(109): ra12.
Moses, A. M., M. E. Liku, et al. (2007). "Regulatory evolution in proteins by turnover and lineage-specific changes of cyclin-dependent kinase consensus sites." Proc Natl Acad Sci U S A **104**(45): 17713-8.
Mustilli, A. C., S. Merlot, et al. (2002). "Arabidopsis OST1 protein kinase mediates the regulation of stomatal aperture by abscisic acid and acts upstream of reactive oxygen species production." Plant Cell **14**(12): 3089-99.
Nair, S. K. and S. K. Burley (2003). "X-ray structures of Myc-Max and Mad-Max recognizing DNA. Molecular bases of regulation by proto-oncogenic transcription factors." Cell **112**(2): 193-205.
Nakagami, H., N. Sugiyama, et al. (2010). "Large-scale comparative phosphoproteomics identifies conserved phosphorylation sites in plants." Plant Physiol **153**(3): 1161-74.
Nash, P., X. Tang, et al. (2001). "Multisite phosphorylation of a CDK inhibitor sets a threshold for the onset of DNA replication." Nature **414**(6863): 514-21.
Nesvizhskii, A. I. (2007). "Protein identification by tandem mass spectrometry and sequence database searching." Methods Mol Biol **367**: 87-119.
Newman, J. R., S. Ghaemmaghami, et al. (2006). "Single-cell proteomic analysis of S. cerevisiae reveals the architecture of biological noise." Nature **441**(7095): 840-6.
Notredame, C., D. G. Higgins, et al. (2000). "T-Coffee: A novel method for fast and accurate multiple sequence alignment." J Mol Biol **302**(1): 205-17.
Oheigeartaigh, S. S., D. Armisen, et al. (2011). "Systematic discovery of unannotated genes in 11 yeast species using a database of orthologous genomic segments." BMC Genomics **12**: 377.
Ohno, S. (1970). Evolution by gene duplication. Berlin, Springer.
Olsen, J. V. and M. Mann (2004). "Improved peptide identification in proteomics by two consecutive stages of mass spectrometric fragmentation." Proc Natl Acad Sci U S A **101**(37): 13417-22.
Ong, S. E., B. Blagoev, et al. (2002). "Stable isotope labeling by amino acids in cell culture, SILAC, as a simple and accurate approach to expression proteomics." Mol Cell Proteomics **1**(5): 376-86.
Otto, S. P. (2007). "The evolutionary consequences of polyploidy." Cell **131**(3): 452-62.
Ozsolak, F. and P. M. Milos (2011). "RNA sequencing: advances, challenges and opportunities." Nat Rev Genet **12**(2): 87-98.
Pache, R. A., M. M. Babu, et al. (2009). "Exploiting gene deletion fitness effects in yeast to understand the modular architecture of protein complexes under different growth conditions." BMC Syst Biol **3**: 74.
Papp, B., C. Pal, et al. (2003). "Dosage sensitivity and the evolution of gene families in yeast." Nature **424**(6945): 194-7.
Parker, J. L., A. M. Jones, et al. (2010). "Analysis of the phosphoproteome of the multicellular bacterium Streptomyces coelicolor A3(2) by protein/peptide fractionation, phosphopeptide enrichment and high-accuracy mass spectrometry." Proteomics **10**(13): 2486-97.

Peng, J., D. Schwartz, et al. (2003). "A proteomics approach to understanding protein ubiquitination." Nat Biotechnol **21**(8): 921-6.
Peng, K., P. Radivojac, et al. (2006). "Length-dependent prediction of protein intrinsic disorder." BMC Bioinformatics **7**: 208.
Pereira-Leal, J. B., B. Audit, et al. (2005). "An exponential core in the heart of the yeast protein interaction network." Mol Biol Evol **22**(3): 421-5.
Pray, L. A. (2008). "Eukaryotic Genome Complexity." Nature Education **1**(1).
Prisic, S., S. Dankwa, et al. (2010). "Extensive phosphorylation with overlapping specificity by Mycobacterium tuberculosis serine/threonine protein kinases." Proc Natl Acad Sci U S A **107**(16): 7521-6.
Proost, S., M. Van Bel, et al. (2009). "PLAZA: a comparative genomics resource to study gene and genome evolution in plants." Plant Cell **21**(12): 3718-31.
Ptacek, J., G. Devgan, et al. (2005). "Global analysis of protein phosphorylation in yeast." Nature **438**(7068): 679-84.
Pu, S., J. Wong, et al. (2009). "Up-to-date catalogues of yeast protein complexes." Nucleic acids research **37**(3): 825-31.
Ravichandran, A., N. Sugiyama, et al. (2009). "Ser/Thr/Tyr phosphoproteome analysis of pathogenic and non-pathogenic Pseudomonas species." Proteomics **9**(10): 2764-75.
Reik, W. (2007). "Stability and flexibility of epigenetic gene regulation in mammalian development." Nature **447**(7143): 425-32.
Reinders, J., K. Wagner, et al. (2007). "Profiling phosphoproteins of yeast mitochondria reveals a role of phosphorylation in assembly of the ATP synthase." Mol Cell Proteomics **6**(11): 1896-906.
Reinders, J., R. P. Zahedi, et al. (2006). "Toward the complete yeast mitochondrial proteome: multidimensional separation techniques for mitochondrial proteomics." J Proteome Res **5**(7): 1543-54.
Rice (2003). "In-depth view of structure, activity, and evolution of rice chromosome 10." Science **300**(5625): 1566-9.
Roskoski, R., Jr. (2005). "Src kinase regulation by phosphorylation and dephosphorylation." Biochem Biophys Res Commun **331**(1): 1-14.
Sadygov, R. G., D. Cociorva, et al. (2004). "Large-scale database searching using tandem mass spectra: looking up the answer in the back of the book." Nat Methods **1**(3): 195-202.
Scannell, D. R., K. P. Byrne, et al. (2006). "Multiple rounds of speciation associated with reciprocal gene loss in polyploid yeasts." Nature **440**(7082): 341-5.
Scannell, D. R., A. C. Frank, et al. (2007). "Independent sorting-out of thousands of duplicated gene pairs in two yeast species descended from a whole-genome duplication." Proc Natl Acad Sci U S A **104**(20): 8397-402.
Schmidl, S. R., K. Gronau, et al. (2010). "The phosphoproteome of the minimal bacterium Mycoplasma pneumoniae: analysis of the complete known Ser/Thr kinome suggests the existence of novel kinases." Mol Cell Proteomics **9**(6): 1228-42.
Schweiger, R. and M. Linial (2010). "Cooperativity within proximal phosphorylation sites is revealed from large-scale proteomics data." Biol Direct **5**(1): 6.
Shou, C., N. Bhardwaj, et al. (2011). "Measuring the evolutionary rewiring of biological networks." PLoS Comput Biol **7**(1): e1001050.

Sickmann, A., J. Reinders, et al. (2003). "The proteome of Saccharomyces cerevisiae mitochondria." Proc Natl Acad Sci U S A **100**(23): 13207-12.
Simillion, C., K. Vandepoele, et al. (2002). "The hidden duplication past of Arabidopsis thaliana." Proc Natl Acad Sci U S A **99**(21): 13627-32.
Souciet, J. L., B. Dujon, et al. (2009). "Comparative genomics of protoploid Saccharomycetaceae." Genome Res **19**(10): 1696-709.
Soufi, B., F. Gnad, et al. (2008). "The Ser/Thr/Tyr phosphoproteome of Lactococcus lactis IL1403 reveals multiply phosphorylated proteins." Proteomics **8**(17): 3486-93.
Soufi, B., C. D. Kelstrup, et al. (2009). "Global analysis of the yeast osmotic stress response by quantitative proteomics." Mol Biosyst **5**(11): 1337-46.
Stark, C., T. C. Su, et al. (2010). "PhosphoGRID: a database of experimentally verified in vivo protein phosphorylation sites from the budding yeast Saccharomyces cerevisiae." Database (Oxford) **2010**: bap026.
Steinmetz, L. M., C. Scharfe, et al. (2002). "Systematic screen for human disease genes in yeast." Nat Genet **31**(4): 400-4.
Sun, X., F. Ge, et al. (2009). "Phosphoproteomic analysis reveals the multiple roles of phosphorylation in pathogenic bacterium Streptococcus pneumoniae." J Proteome Res **9**(1): 275-82.
Supek, F., D. T. Madden, et al. (2002). "Sec16p potentiates the action of COPII proteins to bud transport vesicles." J Cell Biol **158**(6): 1029-38.
Taouatas, N., A. F. Altelaar, et al. (2009). "Strong cation exchange-based fractionation of Lys-N-generated peptides facilitates the targeted analysis of post-translational modifications." Mol Cell Proteomics **8**(1): 190-200.
Thelemann, A., F. Petti, et al. (2005). "Phosphotyrosine signaling networks in epidermal growth factor receptor overexpressing squamous carcinoma cells." Mol Cell Proteomics **4**(4): 356-76.
Thingholm, T. E., O. N. Jensen, et al. (2009). "Analytical strategies for phosphoproteomics." Proteomics **9**(6): 1451-68.
Tirosh, I. and N. Barkai (2007). "Comparative analysis indicates regulatory neofunctionalization of yeast duplicates." Genome Biol **8**(4): R50.
Trinidad, J. C., A. Thalhammer, et al. (2008). "Quantitative analysis of synaptic phosphorylation and protein expression." Mol Cell Proteomics **7**(4): 684-96.
Tuskan, G. A., S. Difazio, et al. (2006). "The genome of black cottonwood, Populus trichocarpa (Torr. & Gray)." Science **313**(5793): 1596-604.
Ubersax, J. A. and J. E. Ferrell, Jr. (2007). "Mechanisms of specificity in protein phosphorylation." Nat Rev Mol Cell Biol **8**(7): 530-41.
Vogel, C. and C. Chothia (2006). "Protein family expansions and biological complexity." PLoS Comput Biol **2**(5): e48.
Wagner, M., R. Adamczak, et al. (2005). "Linear regression models for solvent accessibility prediction in proteins." Journal of computational biology : a journal of computational molecular cell biology **12**(3): 355-69.
Wapinski, I., A. Pfeffer, et al. (2007). "Natural history and evolutionary principles of gene duplication in fungi." Nature **449**(7158): 54-61.
Warnmark, A., E. Treuter, et al. (2003). "Activation functions 1 and 2 of nuclear receptors: molecular strategies for transcriptional activation." Mol Endocrinol **17**(10): 1901-9.

Wilson-Grady, J. T., J. Villen, et al. (2008). "Phosphoproteome analysis of fission yeast." J Proteome Res **7**(3): 1088-97.
Wilson, D., V. Charoensawan, et al. (2008). "DBD--taxonomically broad transcription factor predictions: new content and functionality." Nucleic acids research **36**(Database issue): D88-92.
Wilson, D., R. Pethica, et al. (2009). "SUPERFAMILY--sophisticated comparative genomics, data mining, visualization and phylogeny." Nucleic acids research **37**(Database issue): D380-6.
Wolfe, K. H. and D. C. Shields (1997). "Molecular evidence for an ancient duplication of the entire yeast genome." Nature **387**(6634): 708-13.
Won, A. P., J. E. Garbarino, et al. (2011). "Recruitment interactions can override catalytic interactions in determining the functional identity of a protein kinase." Proc Natl Acad Sci U S A **108**(24): 9809-14.
Wu, R., N. Dephoure, et al. (2011). "Correct interpretation of comprehensive phosphorylation dynamics requires normalization by protein expression changes." Mol Cell Proteomics **10**(8): M111 009654.
Xia, K., Z. Fu, et al. (2008). "Impacts of protein-protein interaction domains on organism and network complexity." Genome Res **18**(9): 1500-8.
Yachie, N., R. Saito, et al. (2009). "In silico analysis of phosphoproteome data suggests a rich-get-richer process of phosphosite accumulation over evolution." Mol Cell Proteomics **8**(5): 1061-71.
Yachie, N., R. Saito, et al. (2011). "Integrative features of the yeast phosphoproteome and protein-protein interaction map." PLoS Comput Biol **7**(1): e1001064.
Yates, J. R., C. I. Ruse, et al. (2009). "Proteomics by mass spectrometry: approaches, advances, and applications." Annu Rev Biomed Eng **11**: 49-79.
Yoshida, R., T. Hobo, et al. (2002). "ABA-activated SnRK2 protein kinase is required for dehydration stress signaling in Arabidopsis." Plant Cell Physiol **43**(12): 1473-83.
Yu, J., J. Wang, et al. (2005). "The Genomes of Oryza sativa: a history of duplications." PLoS Biol **3**(2): e38.
Yuan, J. S., D. W. Galbraith, et al. (2008). "Plant systems biology comes of age." Trends Plant Sci **13**(4): 165-71.
Zotenko, E., J. Mestre, et al. (2008). "Why do hubs in the yeast protein interaction network tend to be essential: reexamining the connection between the network topology and essentiality." PLoS Comput Biol **4**(8): e1000140.

7 Curriculum Vita

Personal Information:

First Name: Ying

Last Name: HE

Title: Msc.

Place of Birth: Shenyang, China

Date of Birth: 10 June 1982

Email: heying340@gmail.com

EDUCATION:

2007.11 – 2011.12	PhD in Bioinformatics VIB Department of , University of Plant Systems Biology, Belgium
2005.9 – 2007.7	MSc in Bioinformatics Wageningen University, The Netherlands
2001.9 – 2005.7	BSc in Biotechnology Dalian University of Technology, China

Research Experience:

2007-2011
PhD research on Bioinformatics and Evolutionary Genomics in Plant Systems Biology department of Gent University and VIB, Belgium. Research focused on which parts of TFs underwent dramatic changes during key phases of plant macroevolution. Comparative analysis on the trends that we observed in plants with trends observed in other eukaryotic lineages, supervised by Prof. Yves Van de Peer.

2007
Research project on developing a web server for taxonomic characterization of sequence samples using signature genes in Comparative Genomics Group in CMBI (Nijmegen University, the Netherlands), supervised by Prof. Martijn Huynen.

2005-2007
Research project on data mining of an integrated annotation resource for protein orthology in Bioinformatics Group in Wageningen University, supervised by Prof. Jack Leunissen.

Conferences, Trainings & Courses:

November 2010	5th EMBO Conference: From Functional Genomics to Systems Biology in Heidelberg, Germany
October 2010	
January 2010	International Mammalian Genome Conference in Crete, Greece
	European course on comparative genomics in Lyon, France
December 2009	5th Benelux Bioinformatics Conference in Liege, Belgium (poster presentation)
October 2009	BioMagnet Attraction Pole 2009 in Gent, Belgium (poster presentation)
July 2009	Advanced course on systems modeling in Cachan, France
January 2009	Asia-Pacific Bioinformatics Conference 2009 in Beijing, China (poster presentation)
December 2008	4th Benelux Bioinformatics Conference in Masstricht, Netherlands (poster presentation)
November 2008	4th EMBO Conference: From Functional Genomics to Systems Biology in Heidelberg, Germany
September 2008	
May 2008	VIBes in Biosciences 2008 in Gent, Belgium
2008-2009	SymBioSys/BioMAGNet event in Leuven, Belgium
	Introduction to Simulation Techniques of the ICES-course in Statistics in Gent University, Belgium
2008-2009	Introductory Statistics of the ICES-course in Statistics in Gent University, Belgium

Publications:

Amoutzias, G., **He, Y.**, Lilley, K., Van de Peer, Y., Oliver, S. (2011) Evaluation and properties of the budding yeast phosphoproteome (under revision in **Molecular Cellular Proteomics**).

The Potato Genome Sequencing Consortium (2011) Genome sequence and analysis of the tuber crop potato. **Nature** 475, 189-195.

Bonnet, E., **He, Y.**, Billiau, K., Van de Peer, Y. (2010) TAPIR, a web server for the prediction of plant microRNA targets, including target mimics. **Bioinformatics** 26, 1566-1568.

Amoutzias, G.*, **He, Y.***, Gordon, J., Mossialos, D., Oliver, S., Van de Peer, Y. (2010) Posttranslational regulation impacts the fate of duplicated genes. **Proc Natl Acad Sci** 107, 2967-2971. (*equal

contribution)

Kuzniar, A., Lin, K., **He, Y.**, Nijveen, H., Pongor, S., Leunissen, J. (2009) ProGMap: an integrated annotation resource for protein orthology. Nucleic Acids Res. 37, W428-W434.

Dutilh, B., **He, Y.**, Hekkelman, M., Huynen, M. (2008) Signature, a web server for taxonomic characterization of sequence samples using signature genes. Nucleic Acids Res. 36, W470-W474.

In preparation:

He, Y., Amoutzias, G., Van de Peer, Y. (2011) Evolution of phosphorylation site in eukaryotes (in preparation).
He, Y., Amoutzias, G., Van de Peer, Y. (2011) Comparative analysis of eukaryotic bZIPs transcription factors (in preparation).

Thesis and Proceedings:

He, Y., Amoutzias, G., Van de Peer, Y. (2009) Evolution of protein-protein interaction networks for plant bZIP transcription factors.
Proceedings of the APBC 2009
He, Y. (2007) Application to identify signature genes. Master thesis in Nijmegen University
He, Y. (2007) Data mining and integration of protein family resources. Master thesis in Wageningen University